



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

A COMPARISON OF APPROACHES:
CLASSICAL, GENERALIZABILITY, AND MULTIFACETED RASCH

by

PETER DENTON MACMILLAN



A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Edmonton, Alberta

Fall, 1995



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

THE AUTHOR HAS GRANTED AN
IRREVOCABLE NON-EXCLUSIVE
LICENCE ALLOWING THE NATIONAL
LIBRARY OF CANADA TO
REPRODUCE, LOAN, DISTRIBUTE OR
SELL COPIES OF HIS/HER THESIS BY
ANY MEANS AND IN ANY FORM OR
FORMAT, MAKING THIS THESIS
AVAILABLE TO INTERESTED
PERSONS.

L'AUTEUR A ACCORDE UNE LICENCE
IRREVOCABLE ET NON EXCLUSIVE
PERMETTANT A LA BIBLIOTHEQUE
NATIONALE DU CANADA DE
REPRODUIRE, PRETER, DISTRIBUER
OU VENDRE DES COPIES DE SA
THESE DE QUELQUE MANIERE ET
SOUS QUELQUE FORME QUE CE SOIT
POUR METTRE DES EXEMPLAIRES DE
CETTE THESE A LA DISPOSITION DES
PERSONNE INTERESSEES.

THE AUTHOR RETAINS OWNERSHIP
OF THE COPYRIGHT IN HIS/HER
THESIS. NEITHER THE THESIS NOR
SUBSTANTIAL EXTRACTS FROM IT
MAY BE PRINTED OR OTHERWISE
REPRODUCED WITHOUT HIS/HER
PERMISSION.

L'AUTEUR CONSERVE LA PROPRIETE
DU DROIT D'AUTEUR QUI PROTEGE
SA THESE. NI LA THESE NI DES
EXTRAITS SUBSTANTIELS DE CELLE-
CI NE DOIVENT ETRE IMPRIMES OU
AUTREMENT REPRODUITS SANS SON
AUTORISATION.

ISBN 0-612-06255-4

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Peter Denton MacMillan.

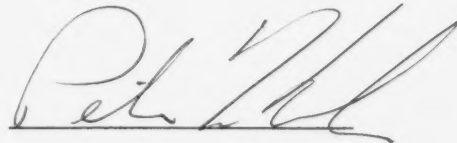
TITLE OF THESIS: A COMPARISON OF APPROACHES:
CLASSICAL, GENERALIZABILITY
AND MULTIFACETED RASCH

DEGREE: Doctor of Philosophy

YEAR DEGREE GRANTED: 1995

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and lend or sell such copies for private, scholarly, or scientific purposes only.

The author reserves all other publication and other rights in association with the copyright in the theses, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



6428 Fairmont Crescent

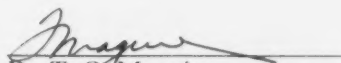
Prince George, B.C.

Canada, V2N 2P5

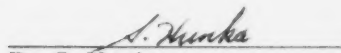
October 4, 1995

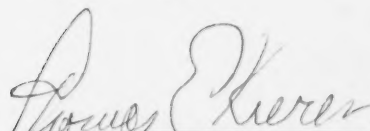
UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES AND RESEARCH

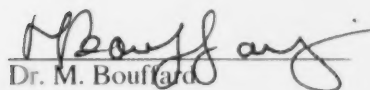
The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled A COMPARISON OF APPROACHES: CLASSICAL, GENERALIZABILITY AND MULTIFACETED RASCH submitted by Peter Denton MacMillan in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

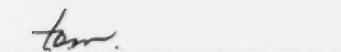

Dr. T. O. Maguire


Dr. W. T. Rogers


Dr. S. Hunka


Dr. T. E. Kieren


Dr. M. Bouffard


Dr. J. O. Anderson

October 2, 1995

DEDICATION

This work is dedicated to

my wife

Mary

my children

Desmond, Fergus, and Eilís

and my parents

Janet E. MacMillan and James A. MacMillan (R.I.P.)

ABSTRACT

The purpose of this study was to compare three well-established approaches, classical (CTT), generalizability (GT), and multifaceted Rasch (MFRM) on their relative abilities to address the problem of rater variability that exists in large scale performance assessments. These approaches were compared on their abilities to detect rater variation in severity, consistency, and agreement. Corrections for rater variability were carried out and compared. The comparisons were made for actual data that consisted of raters' gradings of 4930 examinees' responses to three writing tasks. The examinee responses were marked on a total of 9 scales with each examinee paper graded by 3 raters drawn from a pool of 70 raters. The resulting data matrix was 96% empty.

All three procedures identified rater variation as a problem. The numbers of raters that were identified as different varied greatly between the CTT and Rasch approaches employed. The GT variance component for raters suggested GT agreed with the Rasch All Facet Summary but not with the analytic approaches. However, there was a 90% agreement between CTT and MFRM identification of the 10 most extreme raters but only a 50%-60% agreement when consistency was compared.

The correction procedures for CTT and MFRM produced very similar results. For both procedures a difference between corrected and uncorrected score of one or more points resulted for over 50% of the population. The conditional root mean square was 0.8. When corrections were compared, 75% of the examinees received the same mark for either correction; the correlation between corrections was 1.00 while CRMS was 0.5.

In summary, the practitioner cannot assume that all or any of the three detection approaches will produce similar results. The approach or combination of approaches taken to assess rater variation must be carefully chosen based on the characteristics of interest to

the practitioner. While variation among raters is expected to continue to occur, rater training and monitoring should continue to be carried out. Once the rating has taken place, practitioners should employ a mathematical correction to redress inequities that result due to rater variability.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the many individuals and organizations that made the completion of this dissertation possible. All of the committee members must be thanked for the time and care put into providing me with thoughtful comments and insightful questions. In addition, Dr. Tom Maguire and Dr. Steve Hunka must be thanked for their quiet support and assistance during my time in CRAME. Dr. Todd Rogers has done all this and so much more. His faith in me, his patience with me, his diligence in supervising my work, and his mentorship of me have made it possible for me to complete this degree.

I wish to thank all those at Alberta Education who helped me with the collection of my data. I wish to particularly thank those people within the Analytic section of the Student Evaluation Branch for taking the time to give me a valuable learning experience while I was there. The members of the Kamloops District Teachers' Association must be recognized for their financial assistance. I was provided with one year of paid leave. Without their willingness to support the endeavours of teachers such as myself, I could never have undertaken this program.

My family have supported my academic work for the better part of a decade, first during my Masters, and now during my Doctorate. It is their caring and persistence that is responsible for my ability to carry out this work. Lastly, I am thankful for the many close friends that I have developed as a result of being in CRAME and within Michener Park. The willingness of these people to share both the good times and the hard times has contributed immeasurably to my successful completion of this work.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION	1
The Nature of Variation Among Judges	3
METHODS OF DETECTION AND CORRECTION	6
Classical Test Score Theory	6
Generalizability Theory	8
Item Response Theory	10
COMPARISONS OF METHODS OF DETECTION AND CORRECTION	11
Comparisons of Detection of Rater Variance	12
Comparisons of Correction of Rater Variance	12
THE PROBLEM	14
Selection of the Data Set	16
OUTLINE OF THE THESIS	16
CHAPTER II DESCRIPTION OF DATA	17
THE DATA SET	17
The Choice of the Data Set	17
Appropriateness	17
The Context of the Data	18
Description of the Sections and Scales	19
Rater Selection, Training, and Retraining	22
Selection	22
Training	23
Retraining	24
Marking Procedure	24
THE DATA MATRIX	25
CHAPTER III DISCUSSION OF A PROBLEM AND PROPOSED SOLUTIONS	28
VARIABILITY OF PERFORMANCE ASSESSMENT RATERS	29
Correction Rather than Reduction	31
CLASSICAL TEST THEORY ANALYSIS	32
Monitoring of Rater Variation	32
Resolving Discrepant Ratings	33
Formation of Total Test Score	33

Median versus Mean	34
Unweighted Total Score versus Weighted Total Score	36
Detection of Severity of Rating and the Central Tendency of Ratings	37
Rater severity	37
Central tendency	38
Central Tendency Deviation (CTD).	38
Correction for Differences in Rater Variance.....	39
Linear Scaling.....	39
Scale by scale or total score	41
Linear Regression	41
GENERALIZABILITY THEORY ANALYSIS.....	42
Definition of Variance Components.....	43
Estimation of Variance Components	44
Sampling error.....	46
Detection of Types of Rater Variation.....	46
Desirable size of variance components.....	47
Generalizability and Dependability Coefficients	48
Generalizability coefficients.....	48
Index of dependability coefficients.....	48
The nature of the scale and rater facets.....	49
MULTIFACETED RASCH ANALYSIS	51
From Two Faceted Dichotomous to Multifaceted Polychotomous Models.....	51
IRT Assumptions.....	52
The grading plan in multifaceted Rasch terms.	52
Assumptions	53
One Parameter (Rasch) Item Response Models.....	54
The Dichotomous Rasch Model	54
The Polychotomous Rasch Model.....	54
The Multifaceted Rasch Model.....	55
Estimation of Parameters	56
Multifaceted Rasch Statistics.....	56
Fit of the Data to the Model	57
Outfit.....	57
Infit.....	58
Detection of Types of Rater Variation.....	59
Rater Severity	59

Rater logit severity	60
Chi-square tests.....	60
Rater Consistency	61
Halo.....	62
Rater Agreement.....	62
Rater Group Indices.....	63
Reliability of rater separation.....	63
Rater separation index.	64
Number of strata.....	65
Interrater reliability (IRR)	65
Interpretation of group indices.....	65
Correction for Rater Variance.....	67
Assessment of the size of the correction.....	67
COMPARISON OF DETECTION AND CORRECTION METHODS.....	68
Comparison of Detection of Variation Procedures	68
Comparison of Corrections of Scores	70
Comparisons of Methods of Detection and Correction using Data from a Large Scale Performance Assessment.....	72
Comparisons Made In This Study.....	74
Presentation of Analysis and Results	74
CHAPTER IV THE CLASSICAL APPROACH.....	76
Computer Programs Employed.....	76
PRELIMINARY ANALYSES	76
Psychometric Characteristics.....	76
Weighted versus Unweighted Total Scores	79
MAIN CLASSICAL TEST THEORY ANALYSES	79
Detection of Variation among Rater Severities.....	79
Alexander and Govern Test	80
Results	83
Detection of Central Tendency.....	86
Brown and Forsythe Test	86
Results.....	86
Interrater Reliability	89
Correlations Among Raters	89
Interrater Reliability	90
Coefficient alpha.....	94

Comparisons with other studies.....	95
Correction of Rater Effects.....	95
Linear Scaling.....	95
CHAPTER V GENERALIZABILITY THEORY APPROACH.....	98
Computer Programs Employed.....	98
INITIAL CONSIDERATIONS.....	99
Scales.....	99
Relatively Empty Data Matrix.....	99
PRELIMINARY ANALYSES AND RESULTS.....	100
Sample.....	100
Methods and Results.....	101
Additional preliminary analysis.....	101
MAIN ANALYSES AND RESULTS.....	101
The Sample.....	101
Estimation of Variance Components.....	102
Method and Results.....	102
Interpretation of the Magnitude of Variance Components.....	104
Identification of Raters Who Differ from Other Raters.....	105
Interaction Effects for the Entire Sample.....	109
Decision Studies.....	110
Use of G Study Results Directly in a D Study.....	110
Impact of differing estimates of SEM.....	112
Use of G Study Results to Plan and Assess Alternative D Studies.....	112
Varying the Number of Raters.....	113
Varying the Number of Scales.....	114
An Application of the D Study.....	115
CHAPTER VI MULTIFACETED RASCH MODEL RESULTS.....	117
Computer Programs Employed.....	118
PRELIMINARY ANALYSES AND RESULTS.....	118
Suitability of the Data Matrix.....	118
Assumptions.....	119
Dimensionality.....	119
The Assumption of Equal Discrimination.....	120
The Assumption of Nonspeededness.....	121
Data Misfit Analyses and Results.....	121

Summary of Rasch Preliminary Analyses.....	124
MULTIFACETED RASCH ANALYSES AND RESULTS	124
Rasch All Facet Summary.....	124
Comparisons with other studies.....	127
Rater Characteristics	127
Rater Rasch Severity.....	127
Rater facet severity statistics	129
Rater Consistency	131
Rater Agreement.....	134
Rasch Measurement of Scales	135
Rasch Scale Difficulties.....	135
Scale difficulty statistics	136
Scale Consistency	137
Scale Agreement.....	138
Rasch Measurement of Examinees	139
Examinee Proficiency (Level of Achievement)	139
Examinee facet proficiency.....	139
Examinee Consistency	140
Interpretation of Rasch Point Biserials.....	140
CORRECTION OF EXAMINEE SCORES FOR RATER EFFECTS	141
CHAPTER VII COMPARISONS AMONG APPROACHES.....	143
COMPARISON OF DETECTION APPROACHES.....	143
Rater Severity.....	143
Omnibus Results	144
Identification of Raters Displaying Severity.....	146
Rater Consistency	147
Classical Rater Consistency and Rasch Rater Consistency.....	148
Rater Agreement.....	149
COMPARISON OF CORRECTIONS.....	152
Comparison of Classical and Rasch Corrections	152
A Closing Remark.....	153
CHAPTER VIII CONCLUSIONS.....	155
SUMMARY OF THE STUDY	155
Purpose and Problems.....	155
Data Set.....	155
Analysis and Results.....	156

Preliminary analyses	156
Results of preliminary analyses.....	156
Order of analyses.....	156
Classical Analyses and Results	157
Classical analyses	157
Results of classical analyses	157
Linear scaling correction.....	157
Generalizability Analyses and Results	158
Generalizability analyses.....	158
Generalizability results.....	158
Multifaceted Rasch Analyses and Results	159
Multifaceted Rasch analyses	159
Multifaceted Rasch results.....	159
Rasch correction.....	160
Comparisons.....	160
Comparisons of detection approaches	160
CONCLUSIONS	161
The First Question.....	162
The Second Question	164
LIMITATIONS	164
IMPLICATIONS	165
Implications for Practice.....	165
Implications for Future Research.....	166
REFERENCES.....	168
APPENDIX A JANUARY 1993 MARKER'S MANUAL: ENGLISH 33 DIPLOMA EXAMINATIONS PROGRAM	177
APPENDIX B EQUATIONS FOR CALCULATION OF VARIANCE COMPONENTS.....	194
APPENDIX C DERIVATION OF THE MULTIFACETED RASCH MODEL	196
APPENDIX D WEIGHTED TOTAL MEAN SCORE AND UNWEIGHTED TOTAL MEAN SCORE BY RATER.....	204
APPENDIX E DEMONSTRATION OF THE FEASIBILITY OF THE PROPOSED GENERALIZABILITY SAMPLING DESIGN AND SUBSEQUENT ANALYSIS.....	207

TABLE OF TABLES

TABLE 1	SCALES AND TOTAL EXAMINATION WEIGHTINGS FOR ENGLISH 33 SECTIONS	21
TABLE 2	GENERALIZABILITY AND DEPENDABILITY ERROR TERMS.....	50
TABLE 3	RELIABILITY OF SEPARATION, SEPARATION INDEX, NUMBER OF STRATA, AND INTERRATER RELIABILITY.....	65
TABLE 4	COMPARISONS AMONG APPROACHES.....	75
TABLE 5	MEANS, STANDARD DEVIATIONS, SKEWNESS, AND KURTOSIS OF THE NINE SCALES.....	77
TABLE 6	CORRELATIONS AMONG THE NINE SCALES AND WITH TOTAL SCORES.....	78
TABLE 7	RATER SEVERITIES RANKED BY ALEXANDER-GOVERNZ-SCORE.....	84
TABLE 8	BROWN-FORSYTHE SUMMARY ANOVA	86
TABLE 9	CONFIDENCE INTERVALS FOR RATER CTD	88
TABLE 10	INTERCORRELATIONS BETWEEN A RATER AND ALL OTHER RATERS WHO MARKED A COMMON BUNDLE.....	91
TABLE 11	SPEARMAN-BROWN PREDICTION OF RELIABILITIES.....	94
TABLE 12	FREQUENCY OF SCORE DIFFERENCES.....	96
TABLE 13	THE GENERALIZABILITY GROUP COMPOSITION CHARACTERISTICS	102
TABLE 14	VARIANCE COMPONENTS FOR THE GROUPS AND TOTAL SAMPLE	103
TABLE 15	VARIANCE COMPONENTS FOR TRIPLET 63.....	106
TABLE 16	DEVIATION SCORES FOR EXAMINEES, RATERS, SCALES	107
TABLE 17	DEVIATION SCORES FOR EXAMINEE BY RATER INTERACTIONS.....	108
TABLE 18	DEVIATION SCORES FOR EXAMINEES BY SCALE INTERACTION.....	108
TABLE 19	DEVIATION SCORES FOR RATER BY SCALE INTERACTION	109
TABLE 20	VALUES OF GENERALIZABILITY AND DEPENDABILITY COEFFICIENTS	111
TABLE 21	D STUDY - VARIOUS NUMBERS OF RATERS	113
TABLE 22	D STUDY - VARIOUS NUMBERS OF SCALES	115
TABLE 23	EIGENVALUES FOR ALL NINE FACTORS	120
TABLE 24	MISFITTING RESPONSES.....	123
TABLE 25	RATER RASCH SEVERITY	128
TABLE 26	RATER MEAN FIT STATISTICS	132
TABLE 27	RASCH RATER AGREEMENT	135
TABLE 28	SCALE DIFFICULTIES.....	136
TABLE 29	SCALE MEAN SQUARE FIT STATISTICS	138
TABLE 30	SCALE RASCH POINT BISERIAL.....	138
TABLE 31	EXAMINEE CHARACTERISTICS SUMMARY	139

TABLE 32 DIFFERENCES BETWEEN RASCH FAIR AVERAGES AND OBSERVED SCORES.....	141
TABLE 33 COMPARISON OF SEVERITY DIFFERENCES.....	145
TABLE 34 COMPARISON OF RATER CONSISTENCY.....	148
TABLE 35 COMPARISON OF RATER AGREEMENT	150
TABLE 36 COMPARISON OF CORRECTION PROCEDURES.....	152

TABLE OF FIGURES

1 JANUARY 1993 ENGLISH 33 DIPLOMA EXAMINATION INCLUDING THE WRITTEN RESPONSE SCALES	19
2 REPRESENTATION OF THE DATA MATRIX.	27
3 VARIANCE ESTIMATES FOR COMPONENTS BY GROUP.	104
4 ALL FACET SUMMARY.	125

CHAPTER I INTRODUCTION

There has been a resurgence of interest by the measurement community in performance assessments. While the first and second editions of *Educational Measurement*, released in 1951 and 1971 respectively, devoted entire chapters to the problems posed by performance assessments and essay questions, the third edition of *Educational Measurement*, released in 1989, contains no such chapters. The absence of discussion might be construed as a lack of interest in these topics as the related measurement problems have been solved. However, contrary to this impression, both the 1992 and 1993 National Council on Measurement in Education (NCME) and American Educational Research Association (AERA) Annual Meeting programs included a large number of sessions in which the analyses of performance assessment responses were discussed and debated. The two most common analytical methods employed in the papers presented were generalizability theory and multifaceted Rasch item response models. Interest in these methods has not abated; comparisons and contrasts of the relative merits of these two analyses continue to be made as evidenced by the number of papers presented at the 1994 and 1995 NCME and AERA Annual Meetings.

In spite of the discussions related to performance assessments in the earlier editions of *Educational Measurement*, there is currently a popular assumption found within educational policy debates that "*the new 'authentic' forms of assessment, such as performance assessment* [italics added] and portfolio assessment, are inherently superior to traditional standardized multiple choice tests ... yet little empirical evidence [exists]" to indicate whether or not this assumption is valid (Reardon, Scott, & Verre, 1994, p. 5). "Concerns about subjectivity versus objectivity in evaluating student work" have not been resolved (Darling-Hammond, 1994, p. 7).

The title "performance assessment" suggests that these tests are fundamentally different from other tests that require the use of judges to score the responses. Contrary to this assumption, Fitzpatrick and Morrison (1971) asserted "there is no absolute distinction between performance tests and other tests" (p. 238). In their view, any distinction is merely the degree to which the criterion situation is simulated. They go on to describe performance assessment as an assessment of either a performance or a product.

Physical performances such as operating an airplane flight simulator, taking an automobile driver's road test, and teaching in the classroom are typical performance tasks in which the performance itself is judged. Other tasks are judged by the finished product. Whether the product be a cake, a pair of bookends, or a chemical sample, the quality of the product is judged, and, by implication, the producer of the product is evaluated. The product that is assessed need not be a material product judged on its physical or chemical properties. A journalism student's column in the university newspaper, a mathematics student's problem set, and a graduate student's thesis are all judged on non material aspects of the product. It is the intellectual content of the product as much as the physical product that is of interest. For these students and for many others, the most appropriate task for evaluation purposes is a written response. Each of these written responses is the product of a performance and thus these written responses and all essay writing must be regarded as performance assessment tasks.

While others may debate the claim that written responses in the form of essays should be included as performance assessment data (e.g., Coffman, 1971), there is a clear similarity in the way examinee responses are judged and scored. Stalnaker (1951), speaking directly about essay questions, described the situation as "of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the

subject" (p. 495). Stalnaker's description of essay marking in conjunction with Stiggins (1991) description of performance assessment, "In performance assessment, evaluations of student achievement are based on the professional judgment of the assessor" (p. 265), provides a clear indication that writing assessment should be viewed as performance assessment. Adding further support for the view that essays are performance assessment tasks, Aschbacher (1991), of the Center on Evaluation, Standards, and Student Testing (CRESST), reported that CRESST defines performance assessment to include "direct writing assessments, open-ended questions, hands-on experiments, performances or exhibits, and portfolios of work" (p. 277).

These descriptions and definitions clearly link essay questions, performance assessment tasks, and other subjectively scored questions through a common feature. All share the necessity for scoring by judges and the inevitable variation in the responses of the judges¹.

The Nature of Variation Among Judges

Variation among raters is expected. As Zegers (1991) described the situation, "Judges generally will not agree completely, unless the judgment task is trivial" (p. 321). Judges' ratings must necessarily place examinees in categories whether the rating system allows for whole points or tenths of a point to be awarded. Examinee responses on the other hand can be considered to form a continuum. If for example, categories of 2 and 3 are established, then it is plausible that some examinee responses will occur that will be borderline between these categories. Since the raters have only the option of awarding a 2 or a 3, some will award 2s while others will award 3s. Variation may also occur within raters. An essay that is borderline may receive a 2 from a rater in one marking session but receive a 3 from the same rater in another session. All of this variability among raters or within raters may be described as error since it is a function of the rater rather than the

¹The terms rater, judge, and marker will be used interchangeably throughout this dissertation.

examinee. Further, this variability can be described as random error since there is no predictability as to whether a 2 or a 3 would be awarded in the situations like those just described.

There are several other sources of judge or rater variation that are not random. Some variability among raters is the result of bias, that is, systematic error. Alliger and Williams (1992) described these errors in the following way:

Rating scale 'errors', such as leniency, central tendency, halo ... are recognized as pervasive problems in many areas of applied psychology.... Considerable time and effort have been expended on trying to control or reduce the effects of these rater tendencies. It cannot be said that these efforts have met with much success. (p. 337)

Rater leniency is a tendency to overrate phenomena, while rater severity is a bias that leads to lower scores than warranted (Popham, 1990, p. 300). The second rater bias that affects awarded scores is central tendency bias. Central tendency bias is the preference for marks in the middle of the scale (Popham, 1990, p. 300). A third rater bias that will influence rater variability is the halo effect. A halo effect occurs when a rater uses a general impression of the examinee or the product to affect the ratings given on a variety of separate characteristics (Gronlund & Linn, 1990, p. 225; Popham, 1990, p. 301). The halo effect may occur because of rater bias toward or against a feature of the product itself, for example, a messy handwritten essay versus a laser printed copy on high quality paper. The bias may be from an external source if characteristics of the examinee are known to the rater, for example, ethnic group, gender, past work habits, or attitude. While Gronlund and Linn (1990) and Popham (1990) discuss halo in terms of an interaction between the rater and the examinee's paper, Engelhard Jr. (1994) describes halo as occurring when raters score holistically rather than analytically. Context effects, not discussed by Alliger and Williams (1992), make up a fourth, more elusive bias. Context effects are the result of a reaction to the preceding performances (Hughes and Keeling, 1984). Typically a performance is rated higher by a rater if preceded by a poor

performance and lower if preceded by a good performance. As neither halo effects nor context effects are constant across examinees or procedures, unlike the first two biases, it is not possible to mathematically correct for these forms of bias.

While the goal of training is to develop a pool of interchangeable raters, this goal is rarely if ever achieved (Welch & Miller, 1995, p. 3). The very fact that multiple raters are required to score a performance task is an admission that rater variation, both unsystematic and systematic, does occur. Furthermore, this "variation among raters increases with increased complexity" (Miller & Legg, 1993, p. 11). Unfortunately, rating designs in which not all raters rate all examinees are common as "practical constraints usually prevent the use of large numbers of raters" (Raymond, Webb, & Houston, 1991, p. 102). In this situation, as Guilford realized 40 years earlier, "Ratees [examinees] may benefit or be discriminated against unduly because they happen to be in a certain group" (1954, p. 289). "The random assignment of raters to candidates [examinees] can only decrease, not eliminate, the bias" (Raymond, Webb, & Houston, 1991, p. 102).

It appears rater variation cannot be directly eliminated. Some even fear sacrificing validity for reliability in attempts to do so. However, mathematical adjustments of the rating scores awarded by judges may offer an alternative. Houston, Raymond, and Svec (1991), aware of these procedures and the problems associated with them, still suggested that "because performance ratings are both pervasive and generally unreliable, any method that shows some promise in improving the quality of rating data is worthy of careful consideration" (p. 420). According to Braun (1988) "a systematic study of the effects of different kinds of calibration has yet to be carried out" (p. 3); later, Lunz and Stahl (1990) suggested that corrections have not been routinely applied in the past "because reasonable tools for dealing with the problem were not available" (p. 443). As different methods of detection and correction have recently become feasible, the purpose

of this study was to provide an in-depth comparison of the methods of detection of rater variation and the correction for this variation when these methods are applied to a data set consisting of large scale performance assessment ratings.

Methods of Detection and Correction

There are three main approaches to the detection and correction of the variation among judgments made by raters who score examinee responses to a performance task. These are: the classical test score theory reliability approach, the generalizability theory approach, and the multifaceted Rasch approach. The theory behind each of these approaches is briefly outlined below; a more detailed discussion is presented in Chapter Three.

Classical Test Score Theory

The first approach for detecting variation among raters, and perhaps the most commonly used approach, involves the use of classical test score theory. In classical test score theory (CTT) an examinee's observed score is considered to consist of two components. The first component, a "true score", is defined as the mean observed score over an infinite number of testings for a given examinee. The second component, a "random error", is the discrepancy between the observed score and the true score (Crocker & Algina, 1986, p. 107). Given the postulated random nature of the error, it follows that the variance of observed scores equals the sum of the variance of the true scores and the variance of the random error. The ratio of true score variance to observed score variance is defined as reliability. This reliability coefficient is used to generate a standard error of measurement which in turn allows a confidence band to be constructed around the observed score (Crocker & Algina, 1986).

When examination items² require subjective scoring, variations among raters and the lack of consistency within raters contribute to the error component. Interrater reliability coefficients are typically used to describe the amount of disagreement due to random error. The interrater reliability coefficient, however, is not sensitive to any systematic difference. Therefore a consistent difference between two raters would not be detected with the use of this coefficient.

The reliability coefficients are not the only indices of rater cohesiveness. A second type of index, observer agreement, is "statistically related but conceptually different" (Frick & Semmel, 1978, p. 159). Crocker & Algina (1986) suggest that these "other indices of agreement ... although informative, ... should not be considered substitutes for reliability estimates" (p. 143). Therefore, only reliability estimates will be examined in the present study.

Although a mathematical correction for rater variability could be employed within a classical framework, it is not usually done. Instead the emphasis is placed on the achievement of a satisfactory rater agreement through training, and if needed, retraining procedures. When the satisfactory level of agreement is reached, each examinee receives the mean or median of the scores given by the raters to that examinee. An extra rater may be employed on a case by case basis to replace raters who disagree by an amount that is considered unacceptably large.

While a classical mathematical correction for rater differences, linear scaling, had been predicted to become a useful tool (Guilford, 1954), there appears to be little or no evidence of its use in the literature in the 1980s or 1990s; Nyberg (1987) is one exception. However, regression procedures (de Grujter, 1984; Houston, Raymond, & Svec, 1991; Julian & Searcy, 1995; Raymond & Houston, 1990; Raymond & Roberts,

²For the data in this dissertation, the terms item and scale will be used interchangeably.

1987; Raymond & Viswesvaran, 1993; Raymond, Webb, & Houston, 1991; Wilson, 1988) are often employed to make corrections when the need to do so is shown.

The classical approach, although not a topic of major research interest at the 1992-1995 NCME and AERA conferences, continues to be the method of choice for a number of educational organizations that employ performance tasks as part of their assessment and testing programs. The Student Evaluation Branch of Alberta Education, for example, routinely uses performance questions as part of its provincial school leaving diploma examinations. In scoring these questions, rater variation is controlled through rater training and retraining. Interrater reliability reviews are done, but these serve as an indication of the effect of retraining rather than as a measure of rater reliability. While recommendations suggesting mathematical correction to further control effects of rater variance within this classical approach have been made (Nyberg, 1987), this organization has chosen not to employ them.

Generalizability Theory

A second approach for detecting interrater variation involves the use of generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963). "Generalizability theory has an important role in all forms of educational assessment, including direct writing assessments and performance assessments in other content areas" (Ferrara, 1993, p. 2). "Generalizability theory can be viewed as an extension and liberalization of classical theory that is achieved primarily through the application of analysis of variance procedures to measurement data" (Feldt & Brennan, 1989, p. 127-128). In contrast to classical reliability in which the error is undifferentiated, generalizability theory (G theory) allows the separation of the error component of CTT into multiple sources of error (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In generalizability theory, the term "true score" is often replaced by the term "universe score" to reflect its dependence on the researcher's conception of the universe

of generalization. G theory allows different sources of variability (e.g., raters, examination forms, items, and persons) to all be estimated within one analysis. The data can be analyzed from a relative or absolute frame of reference. Factors can be interpreted as either fixed or random. A generalizability coefficient, analogous to CTT's reliability coefficient, can be calculated. Because of generalizability theory's ability to separate the error into multiple sources, the generalizability coefficient can be defined to match the researcher's conception of universe score. Rater variance, the interaction between rater and examinee, or whatever other source of variance the researcher is able to identify may be specified as part of the true score variance or error variance as appropriate.

While G theory's emphasis is on the estimation of variance components for groups, for example, raters or persons, G theory does produce information about individual raters. For instance, if raters are found to be a large source of variance, then the mean scores of the raters will provide an indication of who the overly lenient or severe raters are. Likewise, if the examinee-rater interaction is large, the appropriate cell means will clarify which combinations of raters and examinees are most troublesome.

G theory has been utilized in a variety of educational performance assessments. It was employed in the British Assessment of Performance Unit's Science study, 1980-1984 (Johnson, 1989), essay writing (Lane and Sabers, 1989), reading, writing, and language skills (Candell & Ercikan, 1992), foreign language speech performance (Kromrey, Bullock, Chason, Du Bose, & Harrison, 1992), science achievement (Baxter, Shavelson, Goldman, & Pine, 1992; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Pine, 1991), portfolio assessment (Koretz, Stecher, Klein, & McCaffrey 1994), and teacher competency (Chauvin, Ellet, Loup, & Naik, 1992).

The information from a generalizability analysis can be used to pinpoint sources of unreliability and produce confidence intervals around observed scores, but it cannot correct for inequities in ratings. Classical mathematical corrections, previously

mentioned, or even IRT solutions could be employed if the generalizability analysis reveals large differences in rater behaviour.

Item Response Theory

The third approach to detecting and correcting rater variation involves the use of a multifaceted one parameter item response model to address the problems of variability in raters and task or item difficulty. The multifaceted one parameter item response model is one of a series of item response models. The collective descriptions of these models is commonly referred to as item response theory (IRT). IRT originally consisted of a two parameter normal ogive model for dichotomous data; the two parameters are item difficulty and item discrimination (Lord, 1952). Currently one-, two-, and three-parameter logistic models for dichotomous data exist (Hambleton, 1989). In the case of the one parameter model only item difficulty is considered, while the three parameter model includes a pseudo-guessing parameter. All three models produce item characteristic(s) for the items and ability estimates for the examinees. The term ability is used in item response models to distinguish the IRT score from the observed score. However, the term "proficiency level" might be a more appropriate term (Hambleton, 1989, p.79).

In addition to the dichotomous models, one and two parameter models for various forms of polychotomous data have been developed (Andrich, 1978; Bock, 1972; Masters 1982; Rasch, 1960 / 1972; Samejima, 1969; 1973). A multifaceted Rasch³ model (MFRM) has been developed by Linacre and Wright (1987). While the first two facets are item difficulty and examinee ability, the other facets are chosen by the researcher much as factors are chosen in generalizability studies. Raters, tasks (groups

³ The one parameter logistic model was developed independently of the two and three parameter models by Rasch (1960 / 1972). This model is commonly referred to as the Rasch model.

of items), occasions, or even gender can be designated as facets if the research questions so dictate.

The problems of detection and correction, treated as separate problems when CTT or G theory are employed, are dealt with simultaneously within the multifaceted Rasch model. The Rasch model will produce an ability estimate corrected for the other facets that were specified in the model. If a model is specified such that rater is a facet, then raters as a group and as individuals are analyzed. It can be quickly seen whether the raters as a group contribute significantly to the differences in ability levels assigned to examinees. At the same time, differences among individual raters are determined. An ability score that accounts for rater differences is then produced. The production of an ability score that is free from task difficulty or rater severity is a feature that has obvious applications for performance assessment data analyses.

The multifaceted Rasch model (Linacre, 1987) has been used for the analysis of performance assessments in a variety of areas including English written expression (Becker, Hess, & Gibney, 1993; Engelhard Jr., 1992; Engelhard Jr., 1994; Hess & Olsen, 1993; Twing & Williams, 1992; Welch & Miller, 1995; Du, 1995), foreign language speaking ability (Kenyon & Stansfield, 1992), public speaking ability (Tatum, 1992), visual arts (Myford, 1992), mathematics problem solving (Brooks & Twing, 1992), and motor skills (Fischer, 1993).

Comparisons of Methods of Detection and Correction

The analyses of interrater variability appear to have coalesced into three camps: the classical approach, the generalizability approach, and the multifaceted Rasch approach. However, little in the way of comparison among the three approaches has been done.

Comparisons of Detection of Rater Variance

The 1993 NCME and AERA conventions contained presentations consisting of comparisons between the ability of G theory analysis and the ability of a multifaceted Rasch analysis to detect rater variation (Marcoulides & Linacre, 1993; Stahl & Lunz 1993). In both presentations data taken from *Generalizability Theory* (Shavelson & Webb, 1991) were analyzed. The comparisons focused on the different emphases that each approach places on various aspects of analysis. Although both presentations gave comparisons of the two approaches, neither data set could be considered representative of real life educational achievement data. Within the generalizability approach, no attempt was made to adjust the observed scores for rater severity. In contrast, Stahl and Lunz provided the observed score and an expected score derived in the Rasch analysis – the "score predicted by the model given the judge's severity and the teacher's ability" (1993, p. 24), but no statistics that summarized the frequency or magnitude of the differences between these scores were provided. This descriptive and sometimes superficial level of analysis has continued to be the norm at the 1994 and 1995 NCME and AERA conventions.

Comparisons of Correction of Rater Variance

A series of three studies that focused on comparing linear regression approaches to correction for rater variation with other methods of correction were carried out in the late 1980s and early 1990s. In these three studies simulated data sets were created and various proportions of the data sets, ranging from 2% to 75%, were eliminated. The correction procedures were then rated on their abilities to recreate the original data set.

In the first study, Raymond and Roberts (1987) compared four procedures: elimination of cases with incomplete data, substitution of the mean for missing data, simple regression, and iterated multiple regression. Three samples with 2%, 6%, and

10% of the data deleted were employed. The two regression procedures outperformed the other procedures; the simple regression performed about as well as the iterated multiple regression (p. 24).

In the second study, Raymond and Houston (1990) compared four mathematical correction methods: ordinary least squares (OLS), weighted least squares (WLS), the multifaceted Rasch model, and data imputation via the E-M algorithm, with the usual procedure of taking the average (or sum) of the ratings across judges. These procedures were applied to a small simulated data set. For this study, 67% of the original data were eliminated in order to create the sample. The results of the corrections were compared to the classical uncorrected estimates. The most important finding of the study was that more accurate estimates of true levels of performance were produced by any of the four methods than when the traditional approach of summing, or averaging, observed ratings was used. Also of interest were the findings that: WLS offered no increase in accuracy over OLS, the correlation between the OLS estimates and Rasch estimates was .986, and the variability of the imputed ratings was quite restricted (p. 15).

In the third study, Houston, Raymond, and Svec (1991) compared OLS, WLS, and imputation employing EM procedures with the usual method of averaging observed ratings. In this study, 50% and 75% of the original data were eliminated in order to create two samples. Again, the three corrections produced more accurate results than the averaging option. In this study the EM impute procedure slightly but consistently outperformed the OLS and WLS procedures. Houston, Raymond, and Svec (1991) cautioned that this relative superiority may not hold if the data vary markedly from multivariate normal. In addition they described the impute procedure as being known for its slow convergence (p. 419). It is noteworthy that in a later non-experimental study in which two of the previous authors took part, when a correction was actually applied, the OLS procedure was employed (Raymond, Webb, & Houston, 1991).

More research needs to be done. Analyses have typically dealt with small numbers of subjects and often a limited number of selected tasks. Often the data were created strictly for the analysis at hand. In contrast, typical large scale performance assessment data sets can be 95% empty and consist of tens of thousands of examinees and in excess of a hundred judges. Comparisons between methods of detection and correction have been relatively few in such cases. Comparisons on typical existing data sets should be done. Many American states are moving to performance assessment at a state-wide level (Aschbacher, 1991). Many Canadian provinces already use performance type tasks in province-wide assessments of students. Alberta Education's Student Evaluation Branch, as an example, routinely rates the responses of tens of thousands of students on a variety of tasks using large numbers of raters. If the use of performance tasks within large scale assessments of performance is again increasing in popularity, then it is reasonable to expect that estimates of performance be made as accurately as is now technically feasible.

The Problem

Given the variety of methods of detection and correction of rater variance employed in performance assessment, and given the relative lack of comparison between these methods, two questions arise:

1. Do Generalizability theory and a multifaceted Rasch analysis produce importantly different indications of rater variability than the indications provided by the classical approach when these analyses are applied to a typical data set produced by students being assessed on a performance assessment task administered on a province-wide basis?

2. Do either a linear scaling, a linear regression, or a multifaceted Rasch analysis produce importantly different indications of examinee ability than those of the uncorrected classical approach when these analyses are applied to a typical data set produced by students assessed on a performance assessment task administered on a province-wide basis?

In the first question "importantly different" will refer to results that would yield different conclusions to be drawn based on the results yielded by the analysis used. Different estimates of the rater variability result in different estimates of the overall reliability of an examination. This is of concern to the examination producer who must have confidence in the instrument and to the examinee who must feel the score is accurate. Different estimates of the rater variability result in different estimates of the size of the confidence intervals to be placed around student scores. This in turn leads to differing decisions made about examinees near a pass or fail cut point. Importantly different also refers to the inability to identify the same raters as aberrant when differing detection methods are employed. Raters who are sufficiently aberrant may be retrained, removed from a rating session, or not allowed to participate in future rating sessions. Essays of those raters may have to be rescored by other raters.

In the second question "importantly different" will refer to results that would yield a one mark variation in a total score of forty five marks. The one mark difference was chosen to maintain a situation consistent with that of an objectively marked (multiple choice) section. In this situation a single flawed item, worth only one point, would not be retained on a provincial examination. It is reasonable that scores for subjectively marked sections should not have a larger tolerance. While it is recognized that a one point multiple choice item and a one point difference on a rating scale may not share the same psychometric meaning, the net worth is the same. If the examinee score is a simple sum of objectively marked and subjectively marked portions of the examination, then the

examinees suffer or benefit equally from a one point change in total score whatever the source of the point.

Selection of the Data Set

The province of Alberta requires its high school students to write province-wide final examinations in a variety of grade 12 subjects. The examinations consist of both objectively scored and subjectively scored portions. It is the subjectively scored portion of an English 33 examination that forms the data set chosen for this study. The particular data set consists of approximately 5000 examinee papers with each paper marked completely by three of the 70 raters employed to score the full set of papers. A more detailed description of the data is provided in Chapter Two, Description of the Data.

Outline of the Thesis

Chapter Two consists of a detailed description of the data set used in this study, including the writing tasks, the data scoring procedures, and the procedures employed to control rater variance and rater bias. Chapter Three begins with a brief discussion of problems involved in essay grading. The approaches to analysis and correction that may be employed in addressing these problems are then described. The major focus of the chapter is on the approaches employed in this study and their relation to the problems in essay grading. Chapter Four contains: a description of the classical analysis, the preliminary analysis, the main analyses, and the results of this analysis. Chapter Five contains the corresponding sections for the generalizability analysis, and Chapter Six, the corresponding sections for the multifaceted Rasch analysis. Chapter Seven consists of a comparison of the results of the three analyses. A summary of the study, conclusions, implications for practice, and implications for future research are presented and discussed in Chapter Eight.

CHAPTER II DESCRIPTION OF DATA

The data used in the present study are described in this chapter. The data consist of judgments by trained raters of examinee responses to the written response portion of the January 1993 English 33 diploma examination. The reasons for the choice of this particular data set are presented first, followed by a detailed description of the written response portion of the examination itself. This is then followed by a description of the rater training procedures and rating procedures employed in the marking of this portion of the examination.

The Data Set

The Choice of the Data Set

Appropriateness

Alberta Education English 33 diploma examination written response data are an appropriate sample for the proposed analyses. Alberta Education reinstated diploma examinations in January 1983. There is a large scale, stable, government funded program for evaluating written response items in place. The procedures employed are likely to reflect accepted theory and common practice employed elsewhere within organizations that assess written response from a classical test score theory perspective. It is recognized that other organizations employing classical analysis techniques may vary in some practices from Alberta Education and from each other, but it is assumed that Alberta Education's procedures are typical of organizations employing classical analysis techniques, and so the results are generalizable to other similar large scale assessments of written response.

The January 1993 English 33 diploma examination data were chosen for this study for several reasons. First, written response has a long history as a performance

task and is still of current interest. Second, the written response portion of the English 33 examination consists of three discrete writing tasks. The written response portion is graded using nine separate but related rating scales. As such, this examination provides a meaningful test of the approaches taken to examine inter-rater reliability. Third, each written response paper is graded in its entirety by three separate raters, thus allowing comparisons among raters and among tasks. Fourth, the English 33 data are typical of a large scale high stakes performance assessment, thereby increasing the likelihood that the findings and the conclusions of this study will be applicable in many other similar situations.

The Context of the Data

The relationship of the English 33 course mark to the English 33 examination marks is outlined in Figure 1. As shown, the English 33 examination total mark accounted for half of the English 33 course mark. In turn, the English 33 examination total mark was formed from two marks, one for the multiple choice portion, the other for the written response portion. It is the written response portion that is of interest in the present study.

As English 33 is a required course designed for students who wish to graduate from high school but who do not plan to enter university, the writing tasks included in the written portion and subsequent scoring of these tasks reflect an emphasis on simple writing competency. As shown in Figure 1, three modes of writing were considered in the January 1993 examination. Each mode resulted in one task, referred to as a section in the examination booklet. The three sections were "Personal Response to Literature", "Functional Writing", and "Response to Visual Communication".

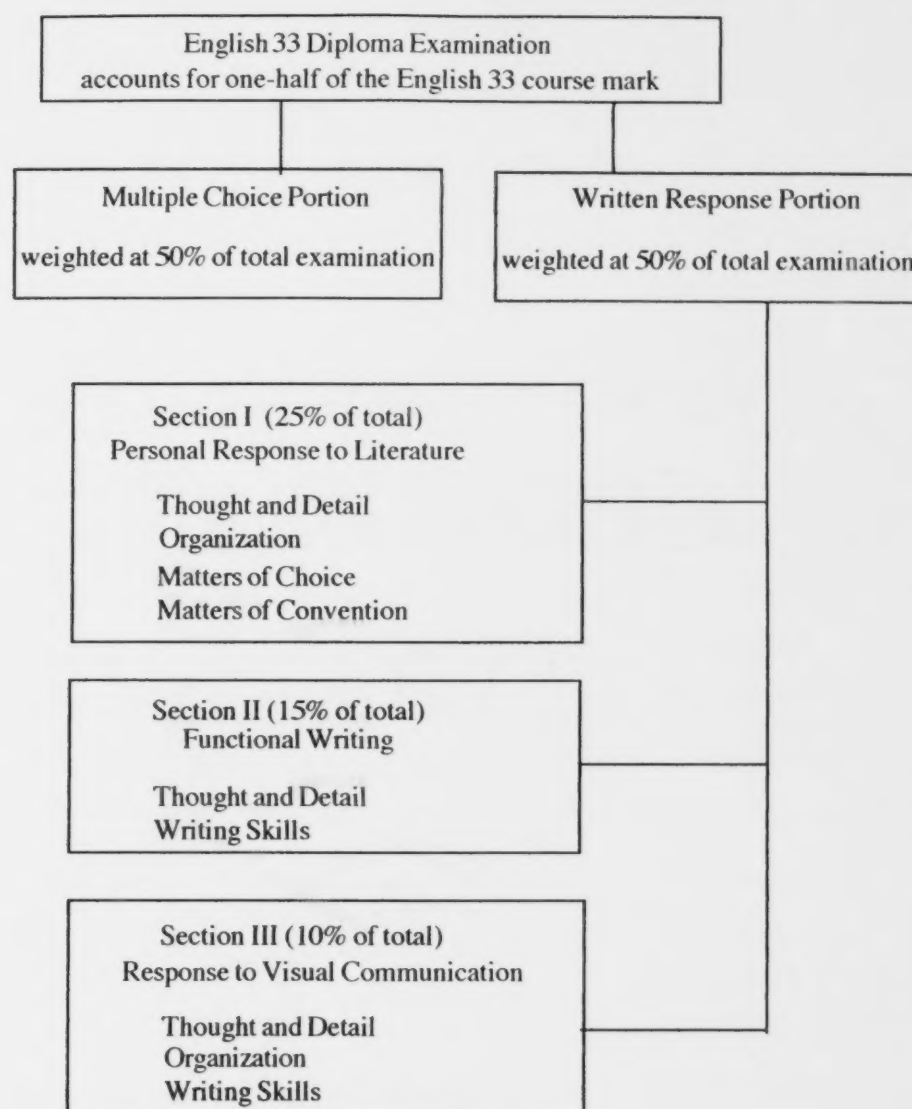


Figure 1. January 1993 English 33 diploma examination including the written response scales.

Description of the Sections and Scales

The first section, Personal Response to Literature, required the examinees to respond to a one page piece of literature given on the examination. The examinees were expected to relate the protagonist's thoughts and feelings to the their own experience.

The second section, Functional Writing, required the examinees to produce a formal letter inviting a past pupil to be a guest speaker during Education Week. The final section required the examinees to respond to a colour reproduction of a photograph. The examinees were to explain what the photograph communicated with their opinions supported by use of the details of the photograph.

The examinees' responses were scored using a series of nine 5-point (1-poor ... 5-excellent) rating scales, presented in Table 1. Each point on each scale, referred to as a scoring category, was defined according to a set of previously established, clearly defined criteria. The *January 1993 Marker's Manual: English 33 Diploma Examinations Program* (Alberta Education, 1993) also prescribed a procedure for cases in which an examinee's response was so scanty or so off topic that the marker judged it to be insufficient for marking (p. 10). In this situation a paper containing the "insufficient" section was referred to a marking supervisor for consideration. If the marking supervisor concurred with the marker, the examinee was awarded a 0 (insufficient). A copy of the portions of the *January 1993 Marker's Manual: English 33 Diploma Examinations Program* (Alberta Education, 1993) that contain section assignments and score category guidelines is located in Appendix A.

As shown in Table 1, the nine 5-point rating scales were weighted to give greater emphasis to some scales than to others. For example, the three Thought and Detail scales accounted for 25 of the possible 50 marks, while the six other scales accounted for the other 25 marks. These 50 marks formed 50% of the total examination score.⁴

⁴The 1994-95 *School Year English 33 Information Bulletin* reports that as of January 1995, the "Organization" scale was dropped from the third section and that the "Writing Skills" scale weighting was increased to 5% (p. 3). The number of raters per paper was reduced to two from three (p. 4).

Table 1
Scales and Total Examination Weightings for English 33 Sections

Section and Scale	Scale Points	Scale Weight
Section I Personal Response to Literature		
Thought and Detail	5	10.0%
Organization	5	5.0%
Matters of Choice	5	5.0%
Matters of Convention	5	5.0%
Section II Functional Writing		
Thought and Detail	5	10.0%
Writing Skills	5	5.0%
Section III Response to Visual Communication		
Thought and Detail	5	5.0%
Organization	5	2.5%
Writing Skills	5	2.5%

In Section I, Personal Response to Literature, the scale topics were Thought and Detail, Organization, Matters of Choice, and Matters of Convention. Two scales, Thought and Detail and Writing Skills were used for Section II, Functional Writing, while Thought and Detail, Organization, and Writing Skills were used for Section III, Response to Visual Communication.

The common name of a scale across sections suggests that the same characteristic is being evaluated across sections. Consider for example, the descriptions for scale point 5 (excellent) for the Thought and Detail scales across the three separate sections (Alberta Education, 1992):

Section I Personal Response to Literature

An insightful understanding of the reading selection(s) is effectively demonstrated. The student's opinion, whether directly stated or implied, is perceptive and is appropriately supported by specific details. Support is precise and thoughtfully selected. (p. 5)

Section II Functional Writing

A precise awareness of audience is effectively sustained. Development of topic or function is clearly focused and effective. Significant information is presented, and this information is enhanced by precise and appropriate details that effectively fulfill the purpose. (p. 9)

Section III Response to Visual Communication

Interpretation of the photograph is insightful and is in the form of an effective generalized idea or theme. Specific details used for support are purposefully chosen and enhance clarity. (p. 11)

As evident in the descriptions for scale point 5 (excellent) across the three sections, different prompts were employed with the three scales that shared a common name. The descriptions for the other four categories of these three Thought and Detail scales paralleled the similarities and differences noted for scale point 5. It is apparent that this scale measured a characteristic that differed across the three sections in spite of the common name, Thought and Detail.

Rater Selection, Training, and Retraining

Selection. The written portion of the English 33 examination was marked by 70 practicing school teachers selected from throughout the province. In order to qualify as a rater, a teacher had to be recommended by his or her superintendent. Further, as required, each selected teacher had taught English 33 for a period of two years, was currently teaching the course, and held a Permanent Professional Certificate. Experience

as a marker, regional representation, and size of student population in the teacher's region were additional factors that were considered when choosing the raters (Alberta Education, 1992, p. 4).

Once selected, teachers were grouped into tables of six with one teacher appointed group leader. Group leaders were experienced markers who attended to minor administrative tasks for their table. They provided some extra advice or guidance but their major role was the marking of papers. As most markers had previous marking experience, the group leaders did not differ greatly from the markers in their groups.

Training. Alberta Education conducted rater training occurred over a two day period immediately preceding the marking period. The first day was spent training group leaders. The training began with reading, but not marking, twelve selected papers bundled in two sets of six. Then the first section of the scoring guide (see Table 1) was introduced, the group leaders' general impressions of student responses to the first section were sought, and the four scoring scales were reviewed. The group leaders then reviewed all of the previously selected example papers on the first section. Following completion of this scoring, each paper was discussed. This process required a full morning. The same sequence was then repeated for the second and third sections during the afternoon. The group leader training concluded with an explanation of organizational details that the group leaders were required to attend to during the marking.

The training of all raters, including the group leaders, took place during the first morning of the following day. The process paralleled that of the group leader training although discussion of the three sections was shortened so that it could be completed in the morning.

The full marking then began in the afternoon. During the marking, as was the case during the training, a copy of all score categories, referred to as a bed sheet, and the marker's manual were kept with each rater at all times.

Retraining. After the first hour of marking, the marking was halted and a "reliability review" was held. The purpose of the reliability review, as described in the marker's manual, was "to promote inter-marker reliability and consistent application of the scoring guides" (p. 17). During the reliability review procedure, all markers were given a common paper that had been selected for its particular characteristics. Usually the reliability review papers were problem papers. These papers were chosen to train markers to correctly apply scale scoring criteria. All raters marked the common paper and gave their marks to the group leader. The group leader then compiled the raters' markings and lead a group discussion in which the reasons for the marks awarded were identified and discussed. After discussion, the markers were invited to alter any of their marks if they wished to. The group then returned to regular marking.

The marks were centrally collected. These marks together with a group summary were posted where all raters could see the results and hopefully draw conclusions that would lead to more uniformity among raters.

Marking Procedure

Prior to the marking session, papers were assembled and put in bundles of six. During marking, a bundle was selected haphazardly by the rater from a designated pickup location, marked, and returned to the designated deposit location where clerks removed the scoring information and returned the bundle to the pickup location. In selecting a bundle the raters checked the bundle for previous marker numbers, if any, so that a bundle was not marked more than once by the same rater. When the bundle had been marked three times the bundle was removed from the marking table.

A total of 4,930 examination papers for the January 1993 English 33 administration were marked by three raters drawn from the pool of 70 raters. Each rater marked an average of 35 bundles during the marking session. This is slightly in excess of 200 papers, although the number of papers each rater actually marked varied from approximately 100 to 400. Given the number of possible rater combinations of three raters chosen from a pool of 70 raters is 54,740, the likelihood of a triplet of raters marking more than one bundle in common is relatively low.

The intact bundles of six were broken in approximately 20% of the cases. The two common breaks resulted in either a five and one split or a three and three split. Reasons for breaks varied: the five-one split occurred when a paper was removed for further scrutiny, while the three-three split occurred near the end of the marking sessions when bundles were split and the two halves given to different raters so that all markers finished the marking at approximately the same time.

Each rater scored all three sections of each paper in a bundle. Although the marker manual is silent on this issue, raters tended to score each paper in its entirety, each section in sequence, before moving onto another paper. Since there was no organized reshuffling of papers within a bundle, papers were likely to be marked in the same order by all three raters within a marking triplet. There appeared to be no discussion of context effects nor any attempt to control for context during the actual marking session.

In contrast, Nyberg (1987) has provided a lengthy discussion of halo effects problems found in her previous analysis of English 33 data. Alberta Education's reliability review papers are carefully selected to remind markers of this effect.

The Data Matrix

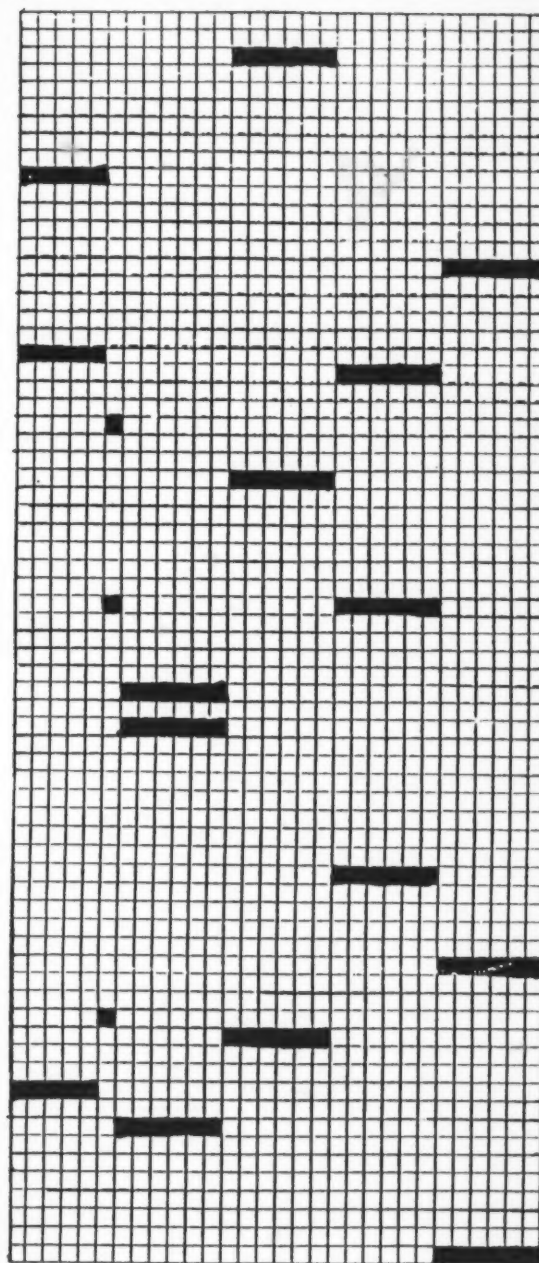
The number of examinees, the use of three raters from a pool of seventy raters, and an examination with nine scales results in a large, $4,930 \times 70 \times 9$, relatively empty

(96% empty) matrix. The rating scores normally range from one to five. As explained earlier (see p. 20), in special cases a series of zeros may be given for an entire section by a rater. This results in the matrix consisting of mainly of scores 1 to 5 with occasional blocks of four, two, or three consecutive 0s.

As a result of the practice of marking papers in bundles, the full matrix can be viewed as made up a series of small sub matrices. In approximately 80% of the cases, six examinees are marked by the same three raters on all nine scales. In the remaining 20%, the sub matrices contain a varying number of examinees, five or three examinees being the most common size for the broken bundles.

Figure 2 illustrates the examinees grouped by bundle. The columns represent examinees while the raters are placed on the rows. To represent all nine scales for the examination, the model would also be nine sheets deep. Reading across any column (examinee) illustrates that an examinee is marked by 3 of the 70 raters. The horizontal bars represent the 6 examinees all marked by a triplet of raters. It is clear that the triplet of raters that marked one bundle were unlikely to mark another bundle as a triplet.

Raters



Examinees

Figure 2.
Representation of the data matrix.

CHAPTER III DISCUSSION OF A PROBLEM AND PROPOSED SOLUTIONS

There are types of examinee responses that require the subjective judgment of a rater at or after the time of response. Experience has revealed that when a common response is rated by more than one rater, variability is found among the raters. This is undesirable as an examinee's score should reflect only the examinee's performance and not the characteristics of those who rated the performance. This is of special concern when examinees are not all rated by the same raters. When this situation occurs, variations due to the raters differentially affect examinee scores. Some examinees will be rated by very severe raters, others by more lenient raters. Some examinees will be rated by very consistent raters, others by more inconsistent raters. Attempts to control for this confounding of the scoring of examinee performance by rater characteristics may employ rater management solutions, mathematical corrections, or a combination of these strategies.

This chapter begins with examples that illustrate that the variance problems associated with the necessary subjectivity of raters have not been eliminated. The examples focus on problems relating to essay questions in particular and performance tasks in general. Next, the methods of detection and mathematical correction of variability of scores due to rater effects are described from within a classical test score framework. Because Alberta Education uses a classical approach to test score analysis, references to Alberta Education practices will be found within this description. Linear scaling and linear regression approaches to correction are then discussed. Following this, a description of a generalizability theory analysis applied to the problem of detection of rater variation is provided. The multifaceted Rasch approach to detection and correction is then presented. Due to its relative newness, the presentation of the multifaceted Rasch analysis includes a discussion of the Rasch statistics that are

employed with this approach. The chapter concludes with a discussion of studies in which the three methods of detection have been compared, and the two methods of correction have been compared.

Variability of Performance Assessment Raters

Problems in scoring performance assessments revolve around rater subjectivity. Given this subjectivity, unwanted variability among raters contributes to error of measurement and low reliability. A review of essay reliability studies, the earliest study cited from the year 1912, the latest from the year 1985, is found in Nyberg (1987, pp. 37-38). One might expect an improvement in reliability with the passage of time, this improvement implying a reduction in rater variability, but there is no evidence of this trend. In Nyberg's (1987) own study of teachers marking Grade 12 English 33 essay examinations, 24 of the 75 raters used in the January 1986 administration had "unsatisfactory correlations, below .8" (p. 118); for the June 1986 marking session, 45 of the 64 raters had unsatisfactory correlations, with some of the interrater correlations as low as .6 (p. 123). Nyberg's findings are of particular interest to this study as both the data for Nyberg's study and the data for this study are samples taken from the marking of the written response section of the English 33 diploma examination. Further, the essay marking procedures described by Nyberg in 1987 have not been altered between that time and the time of the collection of the data employed in the present study (Nyberg, personal communication January 27, 1993), thereby increasing the comparability of the data sets from these three occasions.

Other studies throughout the years 1980 - 1990 also suggest a lack of elimination of rater variation in marking essay questions. For example, Michael, Cooper, Shaffer, and Wallis (1980) conducted a study comparing the essay marking of professors in English departments to the marking of the same essays by professors outside the English department. They found estimated correlations for the composite rating of two raters for

the expert English department readers to range between .65 and .93, while the corresponding correlations for non-English department readers ranged between -.79 and .95. Blok (1985) studied the consistency of rating behaviour between two different occasions. In this study, 16 elementary teachers first marked 105 essays on a scale of 1 to 10 and then repeated the process three months later. The intercorrelations among pairs of scores varied from .42 to .91 (p. 49). Braun (1988) found "a single reading reliability for one essay score to typically be less than .5" when the English Literature and Composition Examination of the Advanced Placement Program data were analyzed (p. 2) and cited Modu and Bleistein (1982) who had obtained results of .47, .54, and .49 for the same examination (p. 13). Findings similar to these were also reported by Littlefield, Harrington, Anthracite, and Garman (1981), Cason and Cason (1984), de Grujter (1984), Lunz and Stahl (1990), and Lunz, Wright, and Linacre (1990).

In apparent contrast to these studies, Hieronymus and Hoover (1987), $r = .91$, and Mullis (1984), $r = .92$, found good agreement. However, in the case described by Mullis (1984) the high interrater reliability was not obtained without a high cost: the use of a combination of five topics and five raters (p. 16). A five topic writing assessment would require an excessive amount of examination writing time. The use of five raters per paper in a large scale assessment would not only be time consuming for the raters involved but also be prohibitively expensive in many situations unless it could be shown to be warranted.

Researchers in the 1990s continue to show evidence of rater variability. Engelhard Jr. (1992) reported an interrater reliability of .82 in his study of a large scale writing assessment (p. 177), while Becker, Hess, and Gibney (1993) reported an interrater reliability of .73 for a large scale writing assessment (p. 12). As Raymond, Webb, and Houston (1991) described the situation, "the literature on the effectiveness of

rater training is mixed" (p. 102). Apparently, attempts to eliminate rater variation continue to be unsuccessful.

Correction Rather than Reduction

The presence of rater variation has been documented; this does not imply attempts to reduce rater variation were not made. Rater variation has been controlled to some degree through the training and management of the raters. The control or reduction of rater variability is seen as a necessary part of the rater training process. Otherwise, "when there is a large amount of discrepancy the value of the judgment procedure will be doubtful" (Zegers, 1991, p. 321). Or, as stated by Engelhard Jr. (cited in Hess & Olsen, 1993), "... if two independent raters provide similar ratings, then this provides support for the quality of the ratings" (p. 9).

Even though the drive toward uniform raters appears reasonable and desirable from a measurement perspective, Huot (1990), in summarizing the influences on rater judgment of writing quality, cited several researchers who express alternate concerns:

Hake (1986) contended that we are losing valuable responses to student writing in the name of interrater reliability. Barritt et al. (1986) also postulated that the need to agree can work against the rater's urge to respond to student writing. Stock and Robinson (1987) stated that the insight available in variant readings is often lost because different readings are treated 'as if they were either incorrect or inaccurate ones'. (p. 105)

As methods of mathematical correction become more feasible, variance among raters is coming to be viewed differently. Lunz and Stahl (1992) have suggested we must "accept that variance among raters ... occurs and must be accounted for before candidate measures are calculated" (p. 17). The solution to rater variation does not lie in the elimination of rater variation but in the correction of scores for rater variation.

Classical Test Theory Analysis

Users of classical test score theory must rely heavily on multiple raters and the management of those raters in order to produce a score that is viewed as accurate and reliable. The amount of error variance is monitored through the calculation of an interrater reliability coefficient, and is reduced through the training of raters.

Monitoring of Rater Variation

One method of monitoring raters is to use a recalibration round in which raters are required to achieve a specified proportion of correct ratings (Meredith & Williams, 1984, p. 13). These recalibration sessions are carried out regularly during the marking session. These sessions, referred to as reliability reviews by Alberta Education, are not reliability studies based on the marking that has taken place during a session, but are sessions in which raters are retrained. First, all raters are asked to mark a common set of papers, then groups of raters compare and discuss their scores. Raters that vary from the group consensus are invited to give their reasons for the scores that they gave. Following this discussion, raters return to the regular marking. The scores from the common review papers are used to calculate an "interrater reliability" coefficient. The use of this process can serve to refocus the raters on the scoring criteria to be followed. However, since raters are aware that the reliability reviews are taking place and that the papers being marked were selected for their usefulness in this retraining exercise, the reliability estimates may not reflect the actual variability among raters during the regular marking.

Other indicators of rater agreement are also employed. Criteria such as percentage of exact score matches or percentage of accurate pass or fail classifications are used in rater training and monitoring (Meredith & Williams, 1984, p. 13). A procedure used by Alberta Education to help control the variability among raters is the daily tabulation of each rater's percent frequency of each scale point awarded along with percent frequency

of each scale point awarded for the entire group of raters. In addition, scale by scale percent frequencies of rater variation from the median scale mark awarded are also calculated and given to each rater on the Individual Marker Daily Summary Sheet (Alberta Education, 1993, p.23). The procedure is intended give raters a sense of whether they are marking too severely or leniently or whether they award the same mark as the median mark for each scale on the papers they have marked in conjunction with two other raters. The percentage of a rater's papers that require rescoring due to scale point discrepancies is also presented to each rater. The Alberta Education belief is that raters will self correct in order to become more like other raters.

Resolving Discrepant Ratings

Procedures that resolve rater differences of greater than one point are frequently used (Meredith & Williams, 1984, p. 14). To illustrate using the Alberta Education situation, if three raters marking a scale produce scores such that one rater differs from both other raters by two or more points, for example (2,2,4) or (2,4,4), the scoring is considered discrepant. In these situations an extra, generally more experienced rater is brought in to rate the paper for that scale. The extra rater is only allowed to assign a mark that falls within the range of marks given by the original three raters. In this example the extra rater would be told that the assigned mark could be 2, 3, or 4. The score given by this fourth rater then becomes the mark for the scale on the paper; the scores of the three initial raters are ignored. Experience at Alberta Education suggests that a fourth reader is required for approximately 5% of the ratings.

Formation of Total Test Score

Once the rating process is complete, the examinee scores are aggregated. Alberta Education uses the examinee's median scale score for the examinee's mark for the scale score. The median scale marks are then weighted (see Table 1) and summed to form a

total score for the written response portion of the examination. The written response portion and the multiple choice portion of the examination are weighted equally and summed to give an examination total score. The combination of multiple choice item scores and weighted median scale scores is employed in the calculation of an estimate of reliability, a coefficient alpha calculated using the SPSS Reliability program (SPSS Reference Guide, 1990, p. 605). Alberta Education carries out the calculation of this coefficient for internal use; its value is not part of any external report.

As described, the variability of the rating process is neglected and the variability of the written portion itself is likely underestimated when computing the total score for the written portion of the examination. Variability among raters on a common paper or variability of raters from occasion to occasion does not enter into the calculation of examination reliability. Consequently it may be that this overestimation of test reliability, and subsequent overestimation of the precision of the examinee score, results in decisions being made that perhaps would not be made given a more accurate estimate of the reliability and standard error of measurement. This overestimation of reliability is at odds with Feldt and Brennan's (1989) advice that "one should gather reliability data in a manner that allows acknowledged error sources to reflect their effects in the intraindividual variation" (p. 107). This inflation of a reliability estimate is not uncommon; Feldt and Brennan (1989) go on to state that due to less than ideal research conditions "reported reliability coefficients tend to overestimate the trustworthiness of educational measures, and standard errors underestimate within-person variability" (p. 108).

Median versus Mean

Within classical test theory, either the median or the mean could be used to summarize the examinee's performance across the judges. Alberta Education uses the median of the three scale scores given by each of the three raters as the score for that

scale. While it may be employed for ease of calculation, the median may also be employed for theoretical reasons. If the categories within a scale are viewed as discrete categories, then a score representing a category in between two categories is undesirable. If an odd number of raters per examinee is used, the median will always be a defined scale score. Mathematically, the median has the advantage over the mean of being resistant to shifts caused by extreme scores by a single rater.

However, the median has the disadvantage of the loss of some information. A median mark of 3 on a five point scale could result from any of the following combinations of marks on three different papers marked by three different raters: (2,3,3), (3,3,3), (3,3,4). A mean mark from the same three raters would be 2.7, 3.0, and 3.3 respectively. If the variation in scores is seen as unwanted variation on papers for which the true score is 3 then the median appears to solve the problem. If the variation in scores is seen as acceptable variation on papers for which the true scores are 2.7, 3.0, and 3.3, then the median hides these differences. The variation would be viewed as acceptable in that categories, while appearing to be discrete, represent a continuum. For example, one might categorize one essay as a 2 or a 3, but it is likely that another essay exists that is of a quality somewhere in between these two categories.

In addition, the median is a terminal statistic. Its usefulness in advanced descriptive and inferential procedures is very limited (Kirk, 1990, p. 117). In contrast, the mean has useful mathematical properties. It is the measure of central tendency that is employed in a variety of interrelated mathematical coefficients such as variances, standard errors, and regression coefficients. The use of the mean would allow comparisons between uncorrected mean scores and the corrected mean scores produced as a result of Hull's linear scaling (Angoff, 1971), linear regression, and the scores produced by the multifaceted Rasch model (Linacre, 1989). The use of the mean so predominates measurement literature that the mean, or a sum of scores, is the only procedure described

in Meredith and Williams' (1984) accounting of identification and control of problems in direct writing assessment. Therefore, to facilitate the comparison made later in this dissertation and with other studies, the mean was used in the present study.

Unweighted Total Score versus Weighted Total Score

While the agreement among raters on individual scales may be of interest to the organization that does the rating, it is the total score that is reported. It is the score upon which decisions are based. Consequently, it is the score of most interest to an examinee. When an examinee's examination performance is summarized by single score, a sum of individual item scores is commonly used. This sum of scores may be simply the addition of all item scores, in which case it is referred to as unweighted. However, a decision may be made to weight a priori the various items. The decision to weight might result for a variety of theoretical or practical reasons. As displayed in Table 1, Alberta Education weighted the scales used for the January 1993 English 33 examination in such a way that the total score was 50 rather than 45, the simple sum of nine 5-point scales.⁵

In the present study, if weighted scores are used, the results will be more easily interpreted by those who wish to judge the results of this study in relation to the actual practices of the organization to whom the data belong. It would be argued that this score is more valid as it is the score that was intended to be used by those who developed the tasks, collected the data, analyzed the data, and interpreted the results. Conversely, if the unweighted score is used, the results could be seen as more generalizable as they are less likely to be due to an artifact of an organization's weighting scheme. However, the possibility exists that the weighted total score may be simply a linear transformation of the unweighted score; in this case the choice of score becomes trivial. Consequently the choice of total score will be made after a set of preliminary analyses designed to

⁵ As of January 1995, the number of scales was reduced to eight. The weighting was adjusted to maintain a total of fifty marks (Alberta Education, 1994, p. 3)

determine this relationship have been carried out; the results of these analyses are presented in Chapter IV.

Detection of Severity of Rating and the Central Tendency of Ratings

Variation among raters may be due to systematic differences among raters in terms of severity of their ratings, their tendency to consistently award the same mark to examinees (the notion of central tendency), and halo. The first two of these unwanted sources of variability among ratings can be detected using classical test score theory. The detection procedures are described next. The third, halo, can be detected using generalizability theory which is an extension of classical test score theory as initially postulated. Consequently, the detection of halo effects will be presented as part of the discussion of generalizability theory.

Rater severity. Using CTT, rater severity is operationally defined as the mean total score taken across examinees. Using this mean, raters are ranked and then compared for severity. Low mean scores are an indication of severity, while high mean scores are an indication of leniency. First, an omnibus test of differences among raters is employed to establish whether significant differences in severity do exist. If significant differences are found, an appropriate multiple comparison test will be selected. Its application will result in the identification of raters that are either too severe or too lenient.

If homogeneity of variance can be assumed, a one way ANOVA would serve as the omnibus test, but as there are varying numbers of papers marked by each rater, " α may be seriously affected" if there is heterogeneity of variances (Glass & Hopkins, 1984, p. 352). Given Nyberg's (1987) study of the English 33 marking procedures in which the 75 raters had standard deviations that ranged from 10.09% to 26.85% (p. 114-115), a procedure that did not require homogeneity of variances was sought. The Alexander and Govern (1994) procedure was selected. In this procedure, a normalized t-score is

employed to produce an omnibus A statistic, which is then used to decide as to whether or not there are significant differences among the raters. If a significant omnibus A is found, the normalized t-score will be used to identify individual raters as being of significantly different severity than the mean rater. The Alexander and Govern (1994) test will be described in detail in Chapter IV.

Central tendency. Popham's (1990) definition of central tendency, the preference for marks in the middle of the scale, if literally applied, would characterize raters according to the proportion of 3s on a series of five point scales that are used to assess the examinee. However, if the intent is to identify raters who gave the same score regardless of examinees, then Popham's (1990) central tendency definition confounds this measure of rater variation with rater severity. To illustrate, a rater who awarded nothing but 3s on a series of five point scales would have the greatest possible central tendency error. Another rater who awarded only 2s would have the same invariance in awarded scores, but because he or she did not use 3s would not be perceived as having a central tendency error. Further to this, both of these two raters would also be identified as possessing halo as defined by Engelhard Jr. (1994). A new definition for central tendency is needed, one that accurately reflects the degree to which a rater consistently awards the same mark, or very similar marks, regardless of that position on the score scale.

Central Tendency Deviation (CTD). In keeping with the notion of central tendency being the consistent use by a rater of that rater's mean mark, a rater's central tendency deviation (CTD_j) was defined in this study as the square root of the mean of the nine scale variances across the examinees marked by a rater:

$$CTD_j = \sqrt{\frac{\sum_{i=1}^9 \sigma_i^2}{9}}, \quad (1)$$

where σ_j^2 is the variance of rater j on scale i . The "9" in the equation reflects the nine scales that were used to rate examinee responses. Values of the CTD_j index smaller than the mean CTD are an indication of a central tendency error. Values of the CTD_j index larger than the mean CTD are also indicative of a systematic error.

Ramsey (1994) suggested the use of two procedures for an omnibus test, with the selection of the procedure dependent on the degree of kurtosis of the distribution of variances. If the distribution is leptokurtic, the Brown and Forsythe (1974) procedure was recommended while the O'Brien (1981) procedure was recommended for normal and platykurtic distributions. The procedure chosen after examination of the distribution of the scores and the application of Ramsey's (1994) recommendation is described in detail in Chapter IV. If the results of this procedure indicates significant differences in CTDs exist, a procedure to produce confidence intervals that allow raters to be classified will be described.

Correction for Differences in Rater Variance

Linear Scaling

Within the classical test score theory framework, Nyberg (1987) suggested a procedure for correcting for rater severity and a procedure for correcting for rater central tendency. The first correction adjusts for rater severity, while the second adjusts for central tendency deviation effects. To correct for rater severity, she recommended that the difference between a rater's mean total score and the group of raters' mean total score be subtracted from each of the individual total scores the rater has given (pp. 141-142). This first correction, a simple addition or subtraction of an amount equal to a known bias was used by Braun (1988) who examined bias corrections for day-of-marking variation as well as rater severity. Nyberg's second suggested correction involves adjusting the

standard deviation of a rater's scores based on the standard deviation of the entire group (Nyberg, 1987, p. 142).

Earlier, Guilford (1954) suggested, "Linear transformations taking care of differences in means as well as differences in standard deviations would become important in this kind of situation [each rater rating only some ratees]" (p. 289). Angoff (1971) credits the process that Guilford described to Hull, 1922, (p. 513). This linear transformation is:

$$\hat{X}_{nj} = M_{CTD} \left(\frac{X_{nj} - M_{X.j}}{CTD_j} \right) + M_{...} \quad (2)$$

where $\hat{X}_{nj}$ is the adjusted total score for examinee n given by rater j,

M_{CTD} is the mean central tendency deviation of the group of raters,

X_{nj} refers to the score given by rater j to examinee n,

$M_{X.j}$ is the mean of rater j,

CTD_j is the central tendency deviation of rater j given in equation (1) and

$M_{...}$ refers to the mean score given by all raters.

With this scaling the examinee's observed score is transformed into a z score for each rater and then transformed according to the characteristics of the group of all raters. Consequently, the adjusted score for an examinee is the score the examinee would receive had he or she been rated by a rater who displayed neither a severity bias nor a CTD bias. The mean of the three adjusted scores taken from the three markings then becomes the examinee's adjusted score, just as the mean of the three unadjusted scores taken from the three markings is taken as the examinee's unadjusted score. This process for adjusting the score is directly analogous to Angoff's (1971) linear equating of examinations. Braun (1988) suggested, "such adjustments are akin to an equating process" (p. 2); in a similar vein, Engelhard Jr. (1994) referred to "an equating model with raters analogous to test forms that may vary in difficulty" (p. 95).

Scale by scale or total score. The scaling described in Equation 2 is to be applied to an examinee's total score. An alternative would be to apply it to each scale. While the scale by scale approach would appear to lead to more accurate results as the rater characteristics would be corrected for each scale, this solution has a source of inaccuracy built into it. The determination of rater characteristics from a mean across all nine scales would give more stable estimates of rater bias for correction purposes. Again drawing on the analogy of equating raters rather than equating test forms, test forms are more commonly equated on the basis of total scores rather than on an item by item basis.

Linear Regression

A second approach within CTT is to use a linear regression correction. A regression model that postulates the observed score consists of a true rating and a severity bias can be expressed by the following equation:

$$X_{nj} = X_n + X_j + \epsilon_{nj} , \quad (3)$$

where X_{nj} is the total score given to examinee n by rater j ,

X_n is the adjusted score for person n ,

X_j is the rater severity correction for rater j , and

ϵ_{nj} is the random error (Raymond & Viswesvaran, 1993, p. 255).

While this particular correction provides examinee scores corrected for rater severity, it does not correct for any differences in rater central tendency deviation bias.

The linear regression transformation can be readily applied on a rater by rater basis when the data matrix is full or relatively full. However, it becomes problematic with a large relatively empty data matrix. The missing-data studies described in Chapter I and that involved a regression solution employed relatively small data sets. While Raymond and Houston (1990) report analyzing large data sets on an IBM mainframe, the

large data set was a 120 x 40 matrix. Other regression studies also employed small data sets: Raymond and Roberts (1987) "sample sizes of 50, 100, 200" (p. 13); Wilson (1988) "20 papers, 19 graders, 76 scores" (p. 80); Houston, Raymond, and Svec (1991) "N=50, N=100" (p. 409); Raymond and Viswesvaran (1993) "40 raters, approximately 115 candidates" (p. 253).

Most recently, Julian and Searcy (1995) managed to apply Raymond and Viswesvaran's (1993) techniques to "a sample of 2678 student papers, each graded by two raters from a pool of 93 operational raters" (p. 10). This sample consisted of 4 four-point scales applied to one two page essay (Julian & Searcy, 1995, p. 8). Julian and Searcy commented that 2770 dummy coded variables were required for the analysis and suggested that "regression models in a fully operational data set may be prohibitive given space and memory boundaries of many computer systems" (p. 15).

The large data matrix considered in the present study – 4930 examinees with 27 scores awarded by 3 raters from a pool of 70 raters – is approximately six times larger than the Julian and Searcy (1995) data set. Its size is such that the use of a linear regression was considered not feasible.

Generalizability Theory Analysis

The appeal of generalizability theory lies in the capability to be able to differentiate various sources of error variance in contrast to classical test score theory where the error is undifferentiated (Smith & Leucht, 1992, p. 229). Miller and Legg (1993) recently commented that "Generalizability theory provides a more complete framework for examining the multiple factors and levels required in alternative [performance] assessment" (p. 11). In a similar vein, Crowley, Thompson, and Worchel (1994) described generalizability theory as subsuming and extending classical test score theory (p. 706).

Generalizability theory has been employed for many years. In 1976, Cardinet, Tourneur, and Allal suggested that times of test administration, scorers, and topics of essay writing are important facets to consider when examining influences upon examinee scores (p. 123). Ferrara (1993) added, "The ... number of ratings necessary to achieve adequate writing score reliability has been a popular facet for generalizability studies" (p. 2). Yet Ferrara (1993) concluded that "generalizability theory and scaling are tools that have not been used often enough in writing assessments" (p. 14). It would seem reasonable to apply generalizability theory to the task of assessing large scale writing assessments. However, large scale assessments do not always lend themselves to routine generalizability analyses. The data set considered in this study is one such case.

Definition of Variance Components

Using generalizability theory notation, the full design of the data set of the present study can be described as a $(n \times j): b \times (i: t)$ (examinees crossed with raters nested within bundles of papers crossed with scales nested within sections) design. However this design is unbalanced in that the number of scales within each section varies across sections and the total number of papers marked by each rater differs. Unbalanced designs present problems in estimation of variance components (Searle, 1971 cited in Shavelson & Webb, 1991, p. 73). "Most computer programs (e.g., BMDP 8V and MIVQUEO in the VARCOMP procedure of SAS) do not have sufficient storage capacities to analyze typical unbalanced designs" (Brennan, Jarjoura, & Deaton, 1980 cited in Shavelson & Webb, 1991, p. 73). A good summary of problems of estimating variance in unbalanced designs is found in Elder (1991).

Shavelson and Webb (1991) suggested that in such situations the design be simplified. There are three approaches: sampling elements to achieve a balanced design, collapsing across levels of a facet, or conducting the analyses within one of the nesting facets. In a discussion of a similar situation involving the decision of whether to average

across facets or report separately, Shavelson and Webb (1991) advised that "the decision to choose between approaches be made primarily on conceptual grounds. The decision maker should decide what kind of information will be needed and analyze the data accordingly" (p. 67).

The first approach is not appropriate since it is not possible to meaningfully sample scales to achieve the same number of scales per section. The second approach is more feasible. The analyses could be completed within each section using a $(j \times n): b \times i$ design. However, with this approach, the design is still unbalanced due to the varying numbers of examinees rated by each rater. Alternatively, and the approach taken here, is to conduct the analysis within each bundle and aggregate the results across bundles. The basic design analyzed is a fully crossed $n \times j \times i$ (examinees-by-raters-by-scale) design.

Estimation of Variance Components

The advantage of employing a fully crossed design is that it yields a complete set of variance components for each facet. It is necessary, though, to take into account whether a facet is fixed or random. If the various members of a facet are considered to be only a sample of many possible members, then a facet is said to be random. If the members of the facet either consist of all possible members of a facet, or if the researcher is only interested in generalizing to those members of the facet, then the facet is said to be fixed (Shavelson & Webb, 1991, pp. 65-66).

If, for example, all facets are considered to be random facets, then the variance components can be determined from the following expressions for the mean squares (MS) yielded in a random effects analysis of variance:

$$EMS_n = n_j n_i \sigma_n^2 + n_i \sigma_{nj}^2 + n_j \sigma_{ni}^2 + \sigma_{nij,e}^2$$

$$EMS_j = n_n n_i \sigma_j^2 + n_i \sigma_{nj}^2 + n_n \sigma_{ji}^2 + \sigma_{nji,e}^2$$

$$EMS_i = n_n n_j \sigma_i^2 + n_n \sigma_{ji}^2 + n_j \sigma_{ni}^2 + \sigma_{nji,e}^2$$

$$\text{EMS}_{nj} = n_i \sigma_{nj}^2 + \sigma_{nji,e}^2$$

$$\text{EMS}_{ni} = n_j \sigma_{ni}^2 + \sigma_{nji,e}^2$$

$$\text{EMS}_{ji} = n_n \sigma_{ji}^2 + \sigma_{nji,e}^2$$

$$\text{EMS}_{nij,e} = \sigma_{nji,e}^2 \text{ (Shavelson \& Webb, 1991, p. 41).}$$

The variance components can then be calculated from these expressions. For example, the variance component for the rater-scale interaction, σ_{ji}^2 , is given by $\sigma_{ji}^2 = (\text{EMS}_{ji} - \sigma_{nji,e}^2) / n_n$. Substitution of the unbiased estimates for the expected mean square estimates $\sigma_{nji,e}^2 = \text{EMS}_{nij,e}$ and EMS_{ji} yields an unbiased estimate for σ_{ji}^2 . This equation and the equations for the remaining variance components are provided in Appendix B.

If, on the other hand, one of the facets is fixed, say scale, then the variance terms σ_n^2 , σ_j^2 , and $\sigma_{nj,e}^2$ will be changed as follows:

$$\sigma_n^{2*} = \sigma_n^2 + \sigma_{ni}^2 / n_i$$

$$\sigma_j^{2*} = \sigma_j^2 + \sigma_{ji}^2 / n_i$$

$$\sigma_{nj,e}^{2*} = \sigma_{nj}^2 + \sigma_{nji,e}^2 / n_i \text{ (Shavelson \& Webb, 1991, p. 68).}$$

As shown the variance components altered by the presence of a fixed facet can be computed from the variance components of a fully random effects analysis.

Sampling error. The estimation of variance components has been called the Achilles heel of generalizability analyses (Shavelson & Webb, 1981, p. 138). One reason is the complexity of some of the designs employed by necessity. A second reason is the variability in the variance component estimate. Sampling errors for variance components have been found to be intolerably large for most practical purposes. Smith (1978) suggested that in order to obtain estimates with acceptable stability, the total

number of observations ($n_n n_j n_i$) needs to be at least 800 (p. 336). However, Lane and Sabers (1989) reported seemingly reasonable results with only 480 total observations.

In the present study, the analysis of one bundle of six papers marked by three judges on nine scales yields only 162 observations. To compensate for this small sample size, the variance component estimates were set equal to the mean of the variance component estimates obtained across replications of bundles. For example, the variance component estimates for raters, σ_j^2 , equaled

$$\sigma_j^2 = \frac{\sum_{b=1}^B \hat{\sigma}_{bj}^2}{B}$$

where $\hat{\sigma}_{bj}^2$ is the variance component estimate for raters within bundle b , and B is the total number of bundles (sub matrices) analyzed. From sampling theory, the error of the mean of these estimates was expected to be small enough to give a variance component estimate with an acceptable level of accuracy.

Detection of Types of Rater Variation

Rater variance, σ_j^2 , is an indicator of rater severity differences. However, there is no variance component calculated from a crossed $n \times j \times i$ design that functions as indicator of rater central tendency. An indicator of halo effect, not identifiable in the CTT analysis, is provided by the interaction between examinees and raters, σ_{nj}^2 . Lastly, rater agreement can be determined within a G theory approach, taking into more full account the nature of the design, the facets and the universe of generalization.

Desirable size of variance components. Desirable outcomes of a fully crossed $n \times j \times i$ G-study analysis would be as follows. The variance due to examinees, σ_n^2 , should be large, accounting for a large percentage of the total variance. This would indicate that the ranking of examinees would be due to reliable differences among the

examinees. In contrast, the rater variance, σ_j^2 , should be small. This would indicate that the raters' variability had little influence on the ranking of examinees, i.e., rater severity was not present. The variance due to scales, σ_i^2 , should also be small. A large variance component for scales would mean the difficulty of scales differed greatly; that is, the amount of proficiency needed to score a 3 on a difficult scale would be greater than the amount of proficiency needed to score a 3 on an easy scale. The implication a large scale variance would be that multiple scales must be used. The interaction between examinees and raters, σ_{ij}^2 , should be small. A large interaction would mean that the examinees were marked differently by different raters, i.e., there were halo effects, as defined by Popham (1990). A large rater-scale interaction, σ_{ji}^2 , would imply scales were marked differently by different raters. Such a finding would suggest that the number of scales should be increased. The interaction between examinees and scales, σ_{ie}^2 , need not be small as a large examinee-scale interaction would indicate that the examinees vary across scales. An examinee who does not display a uniform amount of proficiency on all traits is not a symptom of rater problems. However, where a composite score is used, small scale variance and examinee-scale interaction would support the use of the composite score. The error term, $\sigma_{nji, e}^2$, should be small, indicating that examinee-rater-scale interactions, and random error had little effect on the ranking of the examinees. A large error value might be indicative of misspecification of the model such as the omission of a relevant facet.

Generalizability and Dependability Coefficients

A variety of reliability-like coefficients may be calculated using G theory. If the system being studied only requires the reliable ranking of examinees, then the decision is said to be a relative one; if the examinees are to be reliably sorted into categories, such as pass or fail, then the decision is said to be absolute. Coefficients for relative decisions

are referred to as generalizability coefficients while coefficients for absolute decisions are referred to as dependability coefficients (Shavelson & Webb, 1991, p. 84).

Generalizability coefficients. The general form of the generalizability coefficient is:

$$E_{\rho_{\text{Rel}}^2} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{\text{Rel}}^2},$$

where $E_{\rho_{\text{Rel}}^2}$ is the generalizability coefficient,

σ_r^2 is the universe score variance, and

σ_{Rel}^2 is the variance for the error term for a relative decision.

The variance components that contribute to universe score variance and relative error variance are dependent on the nature of a facet, that is whether a facet is random or fixed.

Index of dependability coefficients. The index of dependability, ϕ , (Brennan & Kane, 1977), is given by the following equation:

$$\phi = \frac{\sigma_r^2 + (\mu - c)^2}{\sigma_r^2 + \sigma_{\text{Abs}}^2},$$

where ϕ is the index of dependability,

σ_r^2 is the universe score variance,

μ is the mean score for the object of measurement,

c is the cut score for the object of measurement, and

σ_{Abs}^2 is the variance for the error term for an absolute decision.

Again, the variance components included within σ_r^2 and σ_{Abs}^2 are dependent on the nature of the facet.

The nature of the scale and rater facets. In the present study, the scale facet was considered to be a random facet. This view was taken for three reasons. First, Alberta

Education would claim that the successful examinees were competent in their use of the English language, not merely that the examinees had a requisite score on a series of nine scales for which there was no meaning beyond those nine scales. Contrary to Shavelson and Webb's (1991) description of the decision that leads to the fixed facet interpretation, the decision maker is interested in generalizing beyond these scales (p. 65). Further, in the Achievement-Over-Time studies conducted by Alberta Education across years in which the number of scales has changed from 10 to 9 to 8, the nature of the generalization has not. This provides evidence that Alberta Education views scale as a random facet. Second, outside agencies interested in the examinees who have passed prerequisite English courses, for example the University of Alberta, are quite willing to accept provincial English 12 examination marks from various provinces as an indication of English competency in spite of different scales being employed for the various examinations in these provinces. Third, a survey of other writing competency studies discussed in the present study reveals a consistent random facet interpretation of scale. Consequently, the scale facet was treated as a random facet.

The raters facet was also considered a random facet. In the case of this facet, the raters that marked any one bundle represented the set of 70 raters. The 70 raters in turn, represented the population of eligible teachers. The number in this population is much greater than 70. Thus raters must also be treated as a random facet.

Presented within Table 2 are the expressions for the relative and absolute variance errors in terms of the variance components for an $n \times j \times i$ (examinee-by-rater-by-scale) design. The definitions of symbols for the variance components have already been given and so will not be repeated. The "n'" notation is used to indicate that the values of n'_n , n'_j , and n'_i may be chosen to be different than the values of n_n , n_j , and n_i in the original generalizability study.

Shown in the first row are the appropriate expressions for a situation in which possible rater variation is ignored. These are included for the purpose of comparison as the relative decision is equivalent to the classical Hoyt's (1941) $n \times i$ (examinees-by-items (scales)) analysis of approach. Shown in the second row are the expressions taking into account rater variation. Thus it would seem that the error terms shown in the second row are appropriate for the design of the data matrix for a single bundle. However, there is another source of unwanted or error variance. While the same 3 raters rate all examinees in a given bundle, raters varied across bundles. Thus the formulas as presented in the second row of Table 2 do not reflect the total error variation. What is missing is the variance source attributable to raters. Incorporating this term leads to the full analysis expressions in the third row of Table 2.

Table 2
Generalizability and Dependability Error Terms

	Generalizability (relative decision)	Dependability (absolute decision)
Hoyt's ANOVA	$\sigma_{Rel}^2 = \frac{\sigma_{nji,e}^2}{n_i}$	$\sigma_{Abs}^2 = \frac{\sigma_i^2}{n_i} + \frac{\sigma_{nji,e}^2}{n_i}$
Bundle analysis	$\sigma_{Rel}^2 = \frac{\sigma_{ni}^2}{n_i} + \frac{\sigma_{nj}^2}{n_j} + \frac{\sigma_{nji,e}^2}{n_i n_j}$	$\sigma_{Abs}^2 = \frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j} + \frac{\sigma_{ij}^2}{n_i n_j} + \frac{\sigma_{ni}^2}{n_i} + \frac{\sigma_{nj}^2}{n_j} + \frac{\sigma_{nji,e}^2}{n_i n_j}$
Full analysis	$\sigma_{Rel}^2 = \frac{\sigma_j^2}{n_j} + \frac{\sigma_{ni}^2}{n_i} + \frac{\sigma_{nj}^2}{n_j} + \frac{\sigma_{nji,e}^2}{n_i n_j}$	$\sigma_{Abs}^2 = \frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j} + \frac{\sigma_{ij}^2}{n_i n_j} + \frac{\sigma_{ni}^2}{n_i} + \frac{\sigma_{nj}^2}{n_j} + \frac{\sigma_{nji,e}^2}{n_i n_j}$

Multifaceted Rasch Analysis

While classical test score theory and its extension, generalizability theory, have been usefully applied to examine achievement and aptitude test performance, there are shortcomings with these models. In particular, classical item difficulties and item

discrimination indices are group-dependent (Hambleton, 1989). Likewise, the person scores are item dependent and when raters are required, the person scores and item difficulties are also dependent on the severity of the raters who rate the examinees. In response to these concerns an alternative framework, item response theory (IRT), was proposed by Lord (1952). Item response theory's item parameters and proficiency parameters are invariant, given that the model fits the data (Hambleton, Swaminathan, & Rogers, 1991, p. 8). IRT applications have been shown to be successful in educational measurement (Hambleton, 1987, p. 132).

From Two Faceted Dichotomous to Multifaceted Polychotomous Models

Early item response models — Lord's (1952) two parameter normal ogive, Rasch's (1960) one parameter logistic, Birnbaum's (1968) two parameter logistic and three parameter logistic (1968) — were all dichotomous models. Recognizing the limitations of dichotomous models, researchers set out to develop polychotomous models. Rasch extended his one parameter logistic model to include a Poisson counts model (Rasch, 1961) and later a binomial trials model (Rasch, 1972). Andrich (1978) and Masters (1982) developed Rasch rating scale and partial credit models respectively. Samejima's (1969) work resulted in the extension of both the two parameter normal ogive model and the two parameter logistic to the ordered category polychotomously scored model. Following this, a nominal response model for unordered categories (Bock, 1972), a continuous response model (Samejima, 1973), and partial credit models (Muraki, 1992) were developed for the two parameter logistic model. These polychotomous models not only allow test items to be scored in ordered polychotomous fashion, but also offer substantial gains in information over the dichotomous counterparts (Donoghue, 1994; Koch, 1983).

All the models just described are two faceted models. They can simultaneously estimate only two parameters. These two parameters are usually examinee proficiency

and item difficulty. The polychotomous models, although improvements over dichotomous models, are still two parameter models. They do not have the ability to simultaneously estimate parameters for examinees, raters, and items on the scores produced in a writing assessment, or in any other rated assessment. In response to this deficiency Linacre (1987) developed what is called the multifaceted Rasch model (MFRM), and the associated computer program, Facets. To date, multifaceted polychotomous two parameter models and multifaceted polychotomous three parameter models do not exist.

Given the relative recency of the multifaceted Rasch model, this section begins with a discussion of assumptions followed by a review of the dichotomous Rasch model, the polychotomous Rasch model, and the multifaceted Rasch model. The description of these models is followed by a description of the statistics commonly employed in a multifaceted Rasch analysis.

IRT Assumptions

The grading plan in multifaceted Rasch terms. Even before the assumptions underlying IRT are tested, the data to be analyzed must be inspected to ensure that sufficient linkage is found among the data to enable estimation of parameters of all members of all facets. This is of particular concern in the present study given the relatively empty data matrix to be analyzed. The selective missing grading plan of the present data set has the necessary linkage for estimation of the Rasch parameters for the various facets (Lunz, 1993). A selective missing grading plan has the following distinguishing characteristics: two or more raters grade each candidate, all raters grade all items, but not all raters grade all candidates. In the present study, 3 raters rated each examinee using a common set of scales, but not all raters graded all examinees.

Assumptions. The data analyzed in this study were collected with the intent of producing one score per examinee. The purpose of this single score was to represent a single trait, an examinee's writing proficiency. Since the score purportedly represents one trait, a unidimensional model was considered. If the assumptions underlying a unidimensional model are met by the data to be analyzed, then the use of a unidimensional model is warranted. Otherwise, a multivariate model would be needed. Assumptions that underlie the unidimensional IRT models are: the items measure only one trait in common and the regression of item scores on latent trait scores follows a normal ogive function or a logistic function. The function is determined by one (difficulty), two (difficulty, discrimination), or three (difficulty, discrimination, pseudo guessing) parameters. In addition to the first two assumptions, it is further assumed that the test in question is non-speeded. The two and one parameter models have the additional assumption of no guessing, while the one parameter model has yet another assumption of equal item discrimination parameters.

As the multifaceted Rasch model employed in this study is a unidimensional model, all the assumptions of the one parameter model were tested. As reported in Chapter VI, the assumptions were met.

One Parameter (Rasch) Item Response Models

The Dichotomous Rasch Model

The original dichotomous Rasch model, expressed in the logarithmic odds on success form, is given by the equation,

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i, \quad (4)$$

where P_{ni} is the probability of examinee n correctly answering item i ,
 $1-P_{ni}$ is the probability of examinee n incorrectly answering item i ,

B_n is the proficiency level of examinee n , and

D_i is the difficulty level for item i .

The original dichotomous model can also be described as a two faceted model with examinee proficiency and item difficulty as the facets.

The Polychotomous Rasch Model

The polychotomous Rasch model may be written as,

$$\ln \left(\frac{P_{nik}}{P_{ni(k-1)}} \right) = B_n - D_i - F_k, \quad (5)$$

where P_{nik} is the probability of examinee n scoring at level k on scale i ,

$P_{ni(k-1)}$ is the probability of examinee n scoring at level $k-1$ on scale i ,

B_n is the proficiency level of examinee n ,

D_i is the difficulty level for scale i , and

F_k is the difficulty of the step from level $k-1$ to level k .

The term, F_k , is equal to the negative of the natural logarithm of the ratio of the probability of an examinee of proficiency 0 scoring k on an item of difficulty 0 to the probability of an examinee of proficiency 0 scoring $k-1$ on an item of difficulty 0. Since the step difficulty is not considered a facet, this polychotomous model is still described as being two faceted in spite of the inclusion of three terms on the right hand side of equation 5.

The dichotomous Rasch model can easily be seen to be a special case of the polychotomous model when the equation for the polychotomous model is rewritten for a dichotomous item:

$$\ln \left(\frac{P_{ni1}}{P_{ni0}} \right) = B_n - D_i - F_1, \quad (6)$$

- where P_{ni1} is the probability of examinee n scoring a 1 rather than a 0 on item i ,
 P_{ni0} is the probability of examinee n scoring a 0 rather than a 1 on item i ,
 that is, $P_{ni0} = 1 - P_{ni1}$,
 B_n is the proficiency level of examinee n ,
 D_i is the difficulty level for item i , and
 F_i is the difficulty level of the step from category 0 to category 1.

The term, F_i , is equal to the negative of the natural logarithm of the ratio of the probability of an examinee of proficiency 0 scoring 1 on an item of difficulty 0 to the probability of an examinee of proficiency 0 scoring 0 on an item of difficulty 0. For a dichotomous item, this ratio is equal to 1. Consequently F_i is equal to 0 and equation 6 reduces to equation 4.

The Multifaceted Rasch Model

A multifaceted Rasch model that includes both a rating scale and a rater severity facet in addition to the scale difficulty facet and the examinee proficiency facet is given by the equation:

$$\ln \left(\frac{P_{nijk}}{P_{nijk-1}} \right) = B_n - D_i - C_j - F_k, \quad (7)$$

- where P_{nijk} is the probability of examinee n being rated at level k by judge j on scale i ,
 P_{nijk-1} is the probability of examinee n being rated at level $k-1$ by judge j
 on scale i ,
 B_n is the proficiency level of examinee n ,
 D_i is the difficulty level for scale i ,
 C_j is the severity of rater j , and
 F_k is the difficulty of the step from category $k-1$ to category k .

A brief description of Linacre's (1989) derivation of this model from the axioms of specific objectivity is provided in Appendix C.

Estimation of Parameters

Derivation of the equations used to obtain unconditional maximum likelihood estimates of the parameters in equation 7 are provided by Linacre (1989). These equations are solved using the Newton-Raphson iteration procedure (Linacre, 1989, p. 95).

Multifaceted Rasch Statistics

A multifaceted Rasch model analysis provides information about individual facet members (examinees, raters, scales) within any facet and information about the facet members as a group. All measures on all members within all facets are placed on one metric, of which the units are logits (Wright & Stone, 1978). In addition, if a rating scale is employed, rating scale points themselves are also placed on the same metric. A variety of statistics are employed to quantify model fit, characteristics of individual facet members, and the facets themselves. A description of these statistics, organized in terms of their use, follows.

Fit of the Data to the Model

Two fit mean square (FitMS) statistics are used to assess the fit of the data to the model before the results of the main analyses are examined. These two related statistics are referred to as outfit and infit statistics. The proportion of misfitting responses to total responses is examined. If this proportion is within an acceptable value, as described below, then the misfitting responses may be examined for relevant patterns that would indicate a misfit for some particular aspect of the model.

Outfit. The outfit statistic is the "usual unweighted mean square residual" with expected value of 1.0 and approximate standard error, $\sqrt{2/df}$, where df is one less than the number of independent replications on which the mean square is based (Lunz, Wright, & Linacre, 1990, p. 336).

For example, the formula for the outfit statistic for rater j is:

$$u_j = \sum_{n=1}^{N_j} \sum_{i=1}^{I_j} \frac{z_{nij}^2}{(N_j + I_j)}, \quad (8)$$

where u_j is the unweighted mean square for rater j ,

z_{nij}^2 is the squared standardized residual for person n on scale i and rater j ,

N_j is the number of persons scored by rater j , and

I_j is the number of scales scored by rater j (Engelhard Jr., 1994, p. 97).

Infit. A disadvantage of the outfit statistic is that it is sensitive to the unexpected responses for a member of the facet under consideration. For example, if the facet of interest is persons in a person-by-item analysis, the outfit statistic is particularly sensitive to unexpected responses made by persons on an item that is far too easy or far too difficult (Wright & Masters, 1982, p. 99). An alternative procedure employs weighted squared residuals so that this undue sensitivity is reduced. Like the outfit, the infit statistic is a "mean square statistic with expectation 1, and range 0 to infinity" but the infit statistic is also an "information-weighted mean-square fit statistic" (Linacre & Wright, 1992, p. 62). The formula for the infit statistic, the weighted mean square, is:

$$v_j = \frac{\sum_{n=1}^{N_j} \sum_{i=1}^{I_j} W_{nij} z_{nij}^2}{\sum_{n=1}^{N_j} \sum_{i=1}^{I_j} W_{nij}}, \quad (9)$$

where v_j is the weighted mean square for rater j ,

z_{nij}^2 is the squared standardized residual, and

W_{nij} is the variance of the observed responses of rater j
(Engelhard Jr., 1994, p. 97).

Values of either the outfit or the infit statistics of much less than one indicate a lack of independence among ratings awarded by a rater, while FitMS values much more than one indicate noise, and / or unmodelled variation (Linacre & Wright, 1992, p. 62).

In order to make comparisons between different values of a fit statistic, the mean squares are standardized to form a t statistic with mean near zero and standard deviation near one (Wright & Masters, 1982, p. 100). This t statistic only approximates the t -distribution. The formula for this statistic, t_o , defined for the outfit statistic for rater j is:

$$t_o = (u_j^3 - 1) \left(\frac{3}{p_j} \right) + \frac{p_j}{3}, \quad (10)$$

where u_j is the unweighted mean square for rater j , and
 p_j is standard deviation of the unweighted mean square.

The infit t_i statistic is defined similarly:

$$t_i = (v_j^3 - 1) \left(\frac{3}{q_j} \right) + \frac{q_j}{3}, \quad (11)$$

where v_j is the weighted mean square for rater j and
 q_j is standard deviation of the weighted mean square.

The t statistic, whether formed from the weighted or unweighted mean square, is seen by some researchers as being excessively sensitive to misfit with relatively large sample sizes (Fischer, 1993, p. 327). Consequently, guidelines have been developed for the mean squares. For example, the interval $0.6 < \text{FitMS} < 1.5$ for acceptable mean square fit has recently found acceptance (Lunz, Wright, & Linacre, 1990, p. 336; Lunz &

Stahl, 1990, p. 433; Engelhard Jr., 1992, p. 176); Hess & Olsen, 1993, p.13). Fischer (1993) employed a similar interval, $0.6 < \text{FitMS} < 1.4$, but added the requirement of the t statistic be in the interval, $-2 < t < +2$ for acceptance of fit (p. 328). Wright and Linacre (1994) more recently suggested the interval, $0.4 < \text{FitMS} < 1.2$, for judged ratings in a system in which judge agreement has been encouraged (p. 370). However, more recently, Smith (1995) cautioned that the use of FitMS guidelines rather than the t statistic guidelines can result in excessive Type I error.

Detection of Types of Rater Variation

Rater Severity

Variation among raters' severities in the multifaceted Rasch analysis is assessed through a variety of statistics. Commonly used statistics include: the logit severities, rater model error, and chi square tests for overall differences among raters (Linacre & Wright, 1992, pp. 59-63). Graphical presentations are also employed. One particularly useful presentation is an all-facet summary graph which allows the easy identification of raters who vary from the main group as well as a comparison of rater severities to relevant characteristics of other facets.

In addition to these indices, reliability-like indices that give specific information about the facet, for example, raters as a group are used. Included in the analysis is the calculation of: a reliability of facet separation estimate analogous to a KR-20, a separation index, and a significance test for differences among individuals. In addition, the number of facet strata, that is, the statistically distinct levels within the facet, can be calculated. An interrater reliability measure has also been suggested. Although these indices are calculated from rater severity values and are used to assess rater severity differences, the use of these indices are reliability-like in nature and so they are discussed under the heading "Rater Group Indices".

Rater logit severity. Severity is the name given to the parameter estimated for raters from the raters' raw scores. The rater logit severity measure, gives an indication of variation among raters, particularly when these differences are compared to the model error. The mean and standard deviation of the logit severities are useful in identifying a rater as being different from other raters.

Chi-square tests. Standard multifaceted Rasch analysis practice produces additional information about raters as a group through the use of two chi-square, χ^2 , tests. The normal chi-square test, with $J-3$ degrees of freedom, tests the hypothesis: the raters can be thought of as a random sample from a normally distributed population (Linacre & Wright, 1992, p. 63). The fixed chi-square test, with $J-1$ degrees of freedom, tests the hypothesis: the raters are all equally severe (Linacre & Wright, 1992, p. 63). The normal chi-square test will yield an insignificant result if the distribution of the rater severities follows a normal distribution. The fixed chi-square test will give a non significant result if all raters are similar.

Rater Consistency

In a multifaceted Rasch analysis, an index describing the consistency of the rater is provided. Consistency refers to the degree to which a rater's awarded scores match the scores predicted for the rater by the multifaceted Rasch model. Consistency is monitored through the infit and outfit statistics.

However, a comparison of the definitions of central tendency and consistency, suggests that the two concepts are similar. Engelhard Jr. (1994) stated, "Central tendency ... as with halo ... introduces an artificial dependency in the rating that leads to overly consistent response patterns that can be detected with rater fit statistics" (p. 99). Yet the respective equations for CTD and infit indicate that these indices are not equivalent. In spite of this non equivalence, empirical evidence exists to support the

claim that low values of the infit statistic can be used to identify central tendency, as defined by Popham (1990). In studies conducted by Engelhard Jr. (1994) and Du (1995), raters that were identified by their low infit and outfit statistics, $< .5$, displayed scoring patterns consistent with the definition of central tendency. For example, one rater in Yi's (1995) study, with an infit statistic of 0.3, awarded the mark pattern "4444" 63% of the time when marking examinees on four 6-point scales. Engelhard Jr. (1994) reported that a rater awarded the scoring pattern "33333" or "22222" 47% of the time when marking examinees on five 4-point scales. The suggested relationship between the seemingly different statistics, CTD_j and rater mean square fit will be examined further in Chapter VII.

Halo. Halo as described by Gronlund and Linn (1990) and Popham (1990), (also see Chapter I), is an interaction between a rater and an individual paper. However, Engelhard Jr. (1994) described the halo effect as being produced when holistic marking was present when analytic was required. He suggested this sequence of scoring would lead to over-consistency, which he described as halo. The examples just presented from the studies of Engelhard Jr. (1994) and Du (1995) were chosen to illustrate the use of the infit statistic in identifying raters that exhibit central tendency as defined by Popham (1990). Other examples reported in these studies demonstrate that the uniform rating patterns need not occur for score points in center of the score point range. Engelhard Jr. (1994) reported that another rater awarded the scoring patterns "4444", "33333", or "22222" 67% of the time when marking examinees on five 4-point scales. He described this pattern as halo. Du (1995) reported two raters with infit of .7 and .6 who awarded "5555" 29% and 33% respectively on a 6-point scale. The rater behaviour, that is, awarding a series of identical marks, may be called central tendency when the marks are 3s on a series of 5-point scales but the term "central tendency" hardly suits if the rater awards a series of 5s on the same 5-point scales.

Rater Agreement

The Rasch rater agreement index that appears most closely related to classical interrater reliability is the Rasch point biserial. The Rasch point biserial is discussed in this subsection.

The formula for the Rasch rater point-biserial is:

$$PB_j = \frac{\sum_{n=1}^{N_j} \sum_{i=1}^{I_j} z_{nji}^2}{\frac{(N_j - 1)(J - 1)(I_j - 1)}{N_j I_j}},$$

where PB_j is the Rasch rater point-biserial ,

z_{nji}^2 is the squared residual scores for judge ratings,

N_j is the number of examinees scored by rater j ,

J is the number of raters, and

I_j is the number of items scored by rater j .

The numerator of the fraction is the squared standardized residual while the denominator is the expression for the degrees of freedom (Wright and Stone, 1979). This point biserial statistic indicates the degree to which a rater grades in the same manner as other raters just as an item point-biserial indicates the degree to which the item functions in the same manner as the other items on the test. There appears to be no previous attempt to establish any minimum acceptable values that would aid in the interpretation of this statistic; the Rasch literature does not address any interpretation of the rater point biserial.

Rater Group Indices

There are a series of indices that describe groups rather than individual members and that are based on the reliability of facet separation index which in turn is based on facet logit measures, for example, rater severities for the rater facet. As such there is a

certain redundancy to these measures. The application of each of these statistics will be described following the presentation of the full set.

Reliability of rater separation. The reliability of separation index, R_j , is an indication of the degree to which the members of a facet can be reliably separated. For raters, a low value of the index is desirable. The formula for the reliability statistic is:

$$R_j = 1 - \frac{\sum_{i=1}^J \frac{S_i^2}{J}}{SD_j^2} ,$$

where R_j is the reliability of separation for raters,

S_j^2 is the rater model error variance,

J is the number of raters, and

SD_j^2 is the observed variance among raters.

The numerator of the fraction is the mean square measurement error, MSE_j (Wright & Masters, 1982, p. 105-106).

Rater separation index. The separation index, G_j , is a measure of the spread of the measures relative to the precision of the measures (Linacre & Wright, 1992, p. 63).

It is calculated using the formula:

$$G_j = \frac{SD_j^2 - \sum_{i=1}^J \frac{S_i^2}{J}}{\sqrt{\sum_{i=1}^J \frac{S_i^2}{J}}} .$$

The numerator of the fraction is referred to as the adjusted standard deviation and the denominator is referred to as the root mean square error, RMSE (Wright & Masters, 1982, p. 106). The separation index is a ratio measure of separation in RMSE units.

The separation index is related to the reliability of rater separation as illustrated by the formulae,

$$R_j = G_j^2 / (1 + G_j^2) \qquad G_j^2 = R_j / (1 - R_j)$$

(Wright & Linacre, 1992, p. 68).

Number of strata. The strata index, H_j , is the number of statistically distinct levels of the measure found within the facet (Wright & Masters, 1982, p. 106). Wright and Masters defined statistically distinct levels as having centers three RMSE apart. If the strata index is calculated for raters, this would indicate the number of distinct levels of raters. The strata index is related to the separation index as illustrated by the formula:

$$H_j = \frac{(4G_j + 1)}{3} ,$$

(Wright & Linacre, 1992, p. 68).

Interrater reliability (IRR) Linacre and Wright (1992) suggest the use of the interrater reliability (IRR) as a measure of the similarity of the elements in the rater facet. An interrater reliability can be calculated by subtracting the reliability of rater separation (R_j), a measure of how different the raters are, from unity; this is expressed in equation form as $IRR = 1 - R_j$.

Interpretation of group indices. As reliability measures are more familiar to researchers than the separation index or strata index, a series of separation index values, strata index values, and IRR values were calculated for given reliability values. These values are reported in Table 3. For example, a reliability of separation of .80 results in a separation index of 2.00 which means the adjusted rater standard deviation is twice the RMSE. For this same reliability of separation, there are 3 strata which implies 3 statistically distinct groups of raters. If, as in the case of English 33, an examinee is rated

by 3 raters drawn from a large pool of raters, it would appear to be ideal if there was only one level of rater severity. Lastly, for this example, the mean interrater reliability is .20, which would suggest very low agreement among raters if this statistic can be interpreted in a classical sense.

Table 3
Reliability of Separation, Separation Index, Number of Strata, and Interrater Reliability

Reliability of Separation	Separation	Strata	IRR
1.00	∞	∞	.00
.95	4.36	6.14	.05
.90	3.00	4.33	.10
.85	2.38	3.51	.15
.80	2.00	3.00	.20
.75	1.73	2.64	.25
.70	1.52	2.36	.30
.60	1.22	1.96	.40
.50	1.00	1.67	.50
.40	0.82	1.42	.60
.30	0.65	1.20	.70
.20	0.50	1.00	.80
.10	0.33	0.77	.90
.00	0.00	0.33	1.00

While a high reliability of separation, near 1.0, may be desirable for examinees, it is not a desirable feature for raters. High reliability of separation for persons is taken as an indication that the person measures obtained indicate differences among the persons. This is not desirable from a rater perspective. An examinee who is exposed to a limited number of raters drawn from a larger pool would be treated fairly only if there was an assurance that the raters all behaved essentially the same. As Table 3 shows, a interrater reliability of .8 or higher indicates the raters all belong to the same stratum. This minimum for *IRR* is the same value as Nyberg's (1987) minimum value for acceptable correlation of a rater with other raters (p. 113).

Correction for Rater Variance

Just as the original item response models produce examinee proficiency estimates that are independent of item effects, a multifaceted Rasch model can produce proficiency estimates that are free from both rater effects and item effects. The scores that are free from influences of the particular items and raters are, in a sense, corrected for the effects that occur when an examinee is exposed to a particular set of items and / or raters. Thus, the item response model that includes a rater facet corrects for rater effects as well as item effects in its production of proficiency estimates.

Assessment of the size of the correction. The corrected score was not easily seen to be corrected in earlier Rasch model analyses. The programs used to complete these analyses only produced proficiency scores measured in logits; there was no routine transformation of the logit score to the observed score metric, hence no ready comparison with the uncorrected score. Linear transformations of the logit scores were always possible, although Wright and Stone's (1979) recommendations lead to a variety of new units, such as *NITS*, *WITS*, and *CHIPS* that are potentially more confusing and awkward than the original logits. The user of the Rasch model, as with other item response models, was left unable to answer the basic question, "How much of a correction was there?"

More recently, the use of a fair average response value, a logit to response-metric conversion has come into use (Linacre and Wright, 1992, p. 61). This fair average "gives the logit measure as an expected average raw response in a standardized environment in which all other elements interacting with this element have a zero logit measure" (Linacre & Wright, 1992, p. 62). This fair average is a linear transformation of the logit score; a linear transformation preserves the interval characteristics of the logit scale (Wright & Stone, 1979, p.192). This response-metric conversion not only allows

the use of a corrected score in the test metric for the individual but now allows the direct comparison between corrected and uncorrected scores. When an observed score and the transformed examinee proficiency score, that is, the fair average multiplied by the number of total examination points are all placed on the same metric, comparisons between the corrected score and uncorrected score can be made.

Comparison of Detection and Correction Methods

A number of selected studies are discussed. The order of presentation is as before, detection of variation followed by correction for that variation. In this section the emphasis is on comparisons among the three approaches. Following this discussion is a description of the comparisons to be carried out in this study.

Comparison of Detection of Variation Procedures

Within CTT, the means and standard deviations of individual raters as well as the correlations between an individual rater and other raters can be calculated. The classical test score model relies on the interrater reliability coefficients to provide measures of error variance attributable to differences among rankings. These interrater reliability coefficients underestimate differences between raters as two raters who consistently disagree by a constant difference will have perfect interrater reliability. The calculation of an internal consistency estimate, for example coefficient alpha, does not take into account rater variability and thus overestimates the reliability when multiple raters are employed.

Generalizability theory, an extension of CTT, offers the ability to produce reliability-like coefficients that more accurately represent the amount of variance associated with the measurements of interest and thus is an improvement over CTT in this situation. Lower, more realistic coefficients are to be expected. A study by Crowley, Thompson, and Worchel (1994) illustrates this feature. Crowley, Thompson, and Worchel (1994) performed a comparison of classical test theory and generalizability

analyses employing an affective instrument, the Children's Depression Inventory (CDI). Both occasion and items were facets for this study. Crowley, Thompson, and Worchel reported internal consistency reliabilities for CDI scores from .86 to .88, coefficients consistent with those previously reported in the literature (p. 710). The generalizability coefficient and dependability index were considerably lower, .63 and .61 respectively (p. 710). These coefficients rose to .81 and .80 for three administrations of the CDI. The "lack of distinction" in CTT reliability estimates for absolute and relative decisions was also highlighted (p. 706). The implications of these coefficients for clinical practice were discussed. Crowley, Thompson, and Worchel gave the case of a child who scored mildly or moderately depressed who was not considered for treatment when examined only once yet was treated after several administrations of the test continued to indicate mild depression.

The multifaceted Rasch model is also capable of identifying both group and individual variation from multiple sources. Both G theory and MFRM have attracted recent interest by the measurement community. This has resulted in two studies involving comparisons of procedures employing Generalizability theory and procedures employing the multifaceted Rasch model. Both were presented at the 1993 joint annual meetings of NCME and AERA (Marcoulides & Linacre, 1993; Stahl & Lunz, 1993). In both studies data were taken from *Generalizability Theory* (Shavelson & Webb, 1991). As these studies were clearly for demonstration purposes, there were serious limitations to both presentations. First, no comparison was made using data from an actual high stakes performance assessment situation. If the usefulness of a procedure cannot be demonstrated under these conditions the procedure is of limited value. Second, no attempt was made to compare any form of corrected observed scores.

By 1995, although some improvement in the quality of comparisons is apparent, the studies continue to demonstrate the same deficiencies. Du (1995) employed only a

small portion of the data available to her. She provided some comparison of raw score and logit score comparisons but continued the MFRM tradition of merely providing a scattergram to demonstrate the lack of linear relationship between the raw and observed scores. No attempt was made to use the MFRM fair average and the raw score mean to allow a comparison of differences expressed in an observed score metric. No detailed Rasch and classical comparisons were made.

Likewise, Schulz and Linacre (1995) employed only a small portion of the available data in their comparison of generalizability and MFRM procedures. The relative ability of these two approaches to deal with larger data sets was skirted. The authors reported a series of reliability coefficients, apparently ignoring Linn and Burton's (1994) advice to the use of SEMs to judge actual differences.

It was the intent of the present study to address these deficiencies. Comparisons were to be made in a common metric so that effects could be judged in light of observable differences. An intact data set was used to be better able to compare the feasibility of these methods under working conditions rather than merely as a researcher's tool.

Comparison of Corrections of Scores

Various mathematical corrections have been suggested: linear scaling solutions, linear regression solutions, and IRT solutions. Linear scaling correction approaches appear to have attracted little recent interest; Braun's (1988) correction for severity being an exception. As described in Chapter I, linear regression correction approaches have been recently championed by a group of researchers revolving around Raymond, for example, Raymond and Houston (1990), Raymond and Roberts (1987), Raymond and Viswesvaran (1993). The linear regression correction approach also has its detractors. Lunz, Wright, and Linacre (1990) argued:

An adjustment for judge severity could be attempted by an analysis of variance (ANOVA) of the raw scores.... But the incomplete data (every judge does not

grade every examination) and the non linearity of the raw scores (they are confined to a finite number of ordered response categories whereas the measures they are meant to imply are not) disqualify this approach. (p. 342)

Rasch measurement practitioners, Lunz and Stahl (1990) also claim as raw scores are not linear, "adjustments made using raw scores are likely to over- or under compensate for differences among judges" (p. 443).

Raymond and Houston (1990) refer to, and counter, Lunz and Stahl's disqualification argument flatly stating that "the models in this paper suggest otherwise" (p. 23). In direct contradiction to the Lunz and Stahl (1990) "linearity" comment, Raymond and Houston (1990) suggest inaccuracies in some Rasch estimates are "most likely due to the logistic transformation, which stretches the tails of the distribution" (p. 15).

Although various mathematical corrections have been recommended and debated, little in the way of empirical comparison among them has been done. The Raymond and Houston (1990) study described earlier compared corrected scores determined using ordinary least squares (OLS), weighted least squares (WLS), and the multifaceted Rasch model. After demonstrating that all of these approaches produced corrected scores that were improvements over the uncorrected scores, Raymond and associated researchers dropped the Rasch model from other comparisons, concentrating on comparisons between the linear regression correction and other non-Rasch models in their subsequent research studies.

The Rasch measurement researchers, as can be seen by the papers discussed in this study, followed a common theme of demonstrating the usefulness of the latest Rasch computer programs in as many varied applications as possible. Comparisons with other IRT models or non-IRT approaches did not appear to interest the Rasch measurement researchers, at least those who presented findings at AERA or NCME Annual General

Meetings during the years 1992 to 1995. Corrected scores in an observed score metric may be given but these were not compared to any scores obtained by other correction procedures.

Comparisons of Methods of Detection and Correction using Data from a Large Scale Performance Assessment

The data set used in the Raymond and Houston (1990) and the Houston, Raymond, and Svec studies (1991) described in Chapter I were simulated data created simply for those studies. Likewise, the Marcoulides and Linacre (1993) and the Stahl and Lunz (1993) papers served to highlight some similarities and differences between Generalizability theory and a multifaceted Rasch model but were not demonstrations of the utility of the procedures under the common measurement conditions of large scale high stakes performance assessment testing.

In contrast to simulated data sets used in many of the other studies, Engelhard Jr. (1992) performed a multifaceted Rasch analysis of a large scale assessment of writing proficiency. Limited percent agreement among raters and interrater reliability information were presented but no attempt was made to compare this information to any of the Rasch analysis information produced by the study. Observed scores corrected by the Rasch analysis were compared to uncorrected observed scores. Unfortunately, Rasch corrected scores were not compared to scores corrected by any other procedure. Engelhard Jr. claimed the "adjustment for rater severity ... improves the objectivity and fairness of the measurement of writing ability" (p. 187). Engelhard Jr. suggested additional research should be done in order "to further examine the Facets [multifaceted Rasch program] model within the context of large-scale assessment of writing ability" (p. 188).

The Alberta Education data used in this study share common features with the Engelhard Jr. (1992) sample. Both sets of data consist of ratings on a series of scales for

a writing sample, ratings by a small number of raters from a large pool of raters, and ratings that are produced only after training has taken place. There are differences as well. The Alberta students were in Grade 12 while Engelhard Jr.'s students, from Georgia, were in grade 8. Nine 5-point scales were used to rate examinee response across three required tasks in Alberta while in Georgia the data consisted of five 4-point scales which were used to rate examinees on one task randomly assigned to them from a total of eight possible tasks. The Alberta tasks match curriculum content areas; it is unlikely that one task per examinee as used in Georgia would do more than simply elicit a written response. While similarities exist, the differences are such that a study should be done that would allow the formation of conclusions generalizable to large scale testing situations similar to that in use in Alberta. To be of interest to a wider audience, comparisons must be made among results obtained by differing approaches, that is, CTT, generalizability theory, and MFRM.

Previous research on an earlier sample of the Alberta data revealed that, like the case in Georgia, the raters varied in severity even after rigorous training and screening. Also, the scales were demonstrated not to be equally reliable (Nyberg, 1987, p. 144). Moreover, the scales that were thought to measure more complex skills were least reliable (Nyberg, 1987, p. 127). As with the Georgia assessment, each paper was scored by a very small number of the large number of possible raters. These similarities suggest important differences between corrected and uncorrected scores will be found with the Alberta sample. Research should be done to determine both the magnitude of these differences and the frequency of these differences.

Differences in rater characteristic have been shown to exist and to persist even after training. These differences will result in differential treatment when only a sample of raters marks any given examination. These differences have been shown to be large enough to significantly affect the scores of the examinees. Corrections have been

suggested. While multifaceted Rasch solutions are in evidence, neither regression corrections nor scaling corrections seem to have been attempted with large scale assessments. The feasibility of these approaches should be investigated. The calls for further research into mathematical corrections for rated data from large scale writing assessment data should be heeded, but this research should be expanded to include comparisons of differing solutions rather than the demonstrations of a particular approach.

Comparisons Made In This Study

There are a variety of indicators within each approach; some are common practice within more than one approach while others are unique to a particular approach. Some indicators provide information primarily about the individual rater differences; some are indicators of group characteristics. A summary of the indicators employed in this study are presented in Table 4. In Table 4, the three approaches are placed in the columns; the rows are indicative of features of each approach that were seen as similar and thus comparable. The order of comparisons in Table 4 follows the general sequence found throughout the study. The rater characteristics are discussed in the order: severity, central tendency, and agreement. Examinee and scale discussions, along with other related measures are included as appropriate. The comparisons conclude with a score correction comparison between the linear scaling correction and the Rasch fair average. The linear regression correction was ruled out as unfeasible for this data set.

Presentation of Analysis and Results

Given the segmented nature of the analyses with each of the three approaches taken, with each step within each segment somewhat dependent upon the results from a preceding step, the analysis and results are presented together in the next three chapters. Chapter IV is concerned with the Classical approach, Chapter V with the Generalizability

approach, and Chapter VI with the multifaceted Rasch approach. The results of the three approaches are compared and discussed in Chapter VII.

Table 4
Comparisons Among Approaches

Classical Test Theory	Generalizability Theory	Multifaceted Rasch Model
Rater Severity Comparisons		
observed scale difficulties	σ_i^2	scale logit difficulties, scale fair averages
Alexander- Govern A statistic rater mean scores including AG z_j	σ_j^2 rater mean scores,	Chi-square test of differences reliability of separation of raters, separation index, strata, rater logit severities, rater fair averages,
Rater Consistency Comparisons		
scale SD		scale infit
BF test of variances CTD,		rater infit and outfit statistics, guidelines for Rasch infit and outfit values
Rater Agreement Comparisons		
scale point-biserial		scale point-biserial
interrater correlations, mean interrater correlation	σ_{nj}^2 , σ_{ji}^2 interactions	rater point-biserial, interrater reliability IRR
coefficient alpha, Hoyt's ANOVA, Spearman-Brown prophecy applied to three raters	σ_j^2 , equivalent G coefficient	reliability of separation for examinees, R_p
Correction Comparisons		
linear scaling correction		fair average

CHAPTER IV THE CLASSICAL APPROACH

The data analyses for the classical approach and the results of these analyses are presented in the present chapter. Prior to the main analyses, three sets of preliminary analyses were carried out. The first was the determination of the psychometric characteristics of the scale scores and the total scores. The second was the examination of the weighted and unweighted test score distributions for the purpose of making a decision as to which score to analyze in this study. The third was the investigation of the feasibility of analysis of section rather than total scores. The main analyses consisted of two parts. In the first part three types of rater characteristics: severities, central tendency deviations, and agreement indices, were investigated. In the second part, a correction in the form of a linear scaling was applied and evaluated.

Computer Programs Employed

The analyses of the classical approach were completed using appropriate procedures chosen from the vast array of programs available through the SPSS-X Data Analysis System, Release 4.0, as documented in the *SPSS Reference Guide* (SPSS Inc., 1990). In addition, Microsoft Excel 4.0, as documented in *Microsoft Excel User's Guide 1* (Microsoft Corporation, 1992), *Microsoft Excel User's Guide 2* (Microsoft Corporation, 1992), and *Microsoft Excel Function Reference* (Microsoft Corporation, 1992) were employed.

Preliminary Analyses

Psychometric Characteristics

The mean, standard deviation, skewness, and kurtosis of the distribution of scores on each of the nine scales and for the weighted and unweighted total scores are reported in Table 5. The corresponding correlations are presented in Table 6. The mean performances across the nine scales ranged from 2.78 (55.6%) to 3.08 (61.6%); the maximum scale score was 5. The standard deviations were less than one score point, varying from .81 (16.2%) to .90 (18.0%). With the exception of the three scales of Section III, the distributions were essentially symmetrical. In the case of the three scales of Section III, the distributions were negatively skewed. The distributions varied in the degree of kurtosis, with four scales showing a fair amount of kurtosis ($\geq .88$). Taken together these findings suggest that in general the scores of the students on the scales were at or above 60% and that the scores were reasonably homogeneous.

Looking now at the total scores, the mean for the unweighted total score was 26.57 (59.0%); the corresponding standard deviation was 5.64 (12.5%). The distribution was very slightly positively skewed and slightly leptokurtic, suggesting that the scores for the majority of the students were within one standard deviation of the mean. The examinee distribution would be expected to have a slight positive skew as the very low achieving students might choose not to write the examination while the very high achiever certainly would write. The weighted score mean and standard deviation results were 29.77 (59.5%) and 6.24 (12.5%) respectively. As with the unweighted total score, the distribution was very slightly positively skewed and slightly leptokurtic. The two distributions were judged to be very similar.

Table 5
Means, Standard Deviations, Skewness, and Kurtosis of the Nine Scales

Section	Name of Scale (Abbreviation)	Mean	SD	Skew	Kurtosis
i	Thought and Detail (persTD)	2.94	0.88	0.02	0.09
I	Organization (persORG)	3.00	0.82	-0.03	0.59
I	Matters of Choice (persMCH)	3.05	0.81	-0.09	0.88
I	Matters of Convention (persMCO)	3.05	0.90	-0.09	0.24
II	Thought and Detail (funcTD)	3.08	0.81	0.03	0.33
II	Writing Skills (funcWRS)	3.01	0.83	-0.03	0.41
III	Thought and Detail (visTD)	2.79	0.85	-0.24	1.03
III	Organization (visORG)	2.78	0.86	-0.18	1.01
III	Writing Skills (visWRS)	2.87	0.88	-0.35	1.15
Total	Unweighted Score	26.57	5.64	0.12	0.57
Total	Weighted Score	29.77	6.24	0.14	0.51

The intercorrelations among scale scores, reported in Table 6, reveal that student performances on the scales were moderately to moderately strongly related, ranging from .32 (persMCO and visTD) to .76 (visTD and visORG). Generally the correlations within a section (writing task) are larger than the between section correlations. Lastly, the uncorrected correlations (row TOT_u) and the corrected correlations (row TOT_c) between each scale score and total unweighted score are reported in the last two rows of Table 6.

As there were only 9 scales, corrected correlations were calculated. In this calculation, each scale was correlated with a total score formed from the other 8 scales; otherwise the correlations would have been spuriously high due to common shared variance between the scale and the total that contained that scale. Both the uncorrected and corrected coefficients were moderate to moderately strong. And, as expected, the corrected coefficients, which ranged from .63 (funcTD) to .80 (visWRS), were higher than the uncorrected correlations which ranged from .53 (funcTD) to .73 (visWRS).

The correlation between the unweighted and weighted total scores (not reported in Table 6) was .98.

Table 6
Correlations Among the Nine Scales and With Total Scores

	Section I (pers)				Section II (func)		Section III (vis)		
	TD	ORG	MCH	MCO	TD	WRS	TD	ORG	WRS
TD	—	.74	.64	.56	.37	.42	.36	.36	.40
ORG		—	.67	.59	.38	.45	.35	.41	.43
MCH			—	.75	.37	.58	.35	.39	.54
MCO				—	.36	.64	.32	.39	.60
TD					—	.58	.36	.37	.40
WRS						—	.37	.43	.63
TD							—	.76	.62
ORG								—	.69
WRS									—
TOT _u	.73	.75	.79	.78	.63	.76	.68	.72	.80
TOT _c	.64	.68	.73	.70	.53	.69	.58	.63	.73

Weighted versus Unweighted Total Scores

As mentioned earlier, a decision was to be made between the use of an unweighted total score or a weighted total score. If, as previously mentioned, the weighted score is simply a linear transformation of the unweighted score, then the choice of score is of no consequence. As reported earlier, when expressed as a percentage, the mean and standard deviation of the unweighted total are essentially the same as the corresponding values for the weighted total (i.e., 59.0% vs. 59.5%; 12.5% vs. 12.5%). Further, as previously reported, the shapes of the two distributions are very similar, and the correlation between the two scores is almost one (i.e., .98).

Given that many of the subsequent analyses were to involve raters, differences between the weighted and unweighted totals and the correlation between these scores were also examined for each rater. These results, reported in Appendix D, revealed that the means expressed as percentages, are essentially the same for each rater. For all raters the correlations between the two total scores exceeded .97.

Taken together, these results indicated that subsequent comparisons among the classical, generalizability, and multifaceted Rasch approaches would not be differentially affected by the use of either of these total scores. Thus for greater simplicity and greater generalizability, the unweighted total score was employed in all subsequent analyses in which a total score was considered. All further references to total score are references to the unweighted total score.

Main Classical Test Theory Analyses

Detection of Variation among Rater Severities

As described in Chapter III, rater severity was defined in terms of the ranked distribution of mean awarded scores by the rater for the sample of papers each rater

marked. In the event that there was homogeneity of rater variances, an ANOVA procedure could be employed as an omnibus test of differing severities. To test for homogeneity of rater variance, the Bartlett-Box test was carried out. The results indicated significant heterogeneity, $F_{69, 14720} = 6.49$, $p < .0005$. Therefore to test for differing levels of severity among raters, the omnibus statistic proposed by Alexander and Govern (1994) was employed.

Alexander and Govern Test

The Alexander and Govern (1994) test was designed for use under conditions of heterogeneity of the variances of raters and has "Type I error rates that are very near nominal and Type II error rates that are close to James's (1951) second-order approximation" (p. 91). Normalized t like scores for this test, AG_{z_j} , are used to compute the Alexander and Govern A statistic,

$$A = \sum_{j=1}^J AG_{z_j}^2,$$

where A is the Alexander and Govern test statistic,

AG_{z_j} is a normalized Student's t statistic, and

J is the total number of raters (Alexander & Govern, 1994, p. 93).

The normalized t scores are computed as follows. First a variance-weighted common mean is used; the weighting factor for rater j is given by

$$w_j = \frac{1/S_j^2}{\sum_{j=1}^J 1/S_j^2},$$

where w_j is the weight to be applied to rater j ,

S_j^2 is the squared standard error of the mean of rater j , and

J is the total number of raters (Alexander & Govern, 1994, p. 92-93).

This variance-weighted common mean is given by

$$M_w = \sum_{j=1}^J w_j M_j$$

where M_j is the rater mean, otherwise known as the rater severity, and M_w is the weighted mean. The one sample t test statistic for rater j , t_j , is computed by,

$$t_j = \frac{M_j - M_w}{S_w}.$$

These t scores are then normalized to yield a distribution AG_{z_j} with mean zero and variance 1. The A statistic is distributed as χ^2 with $J - 1$ degrees of freedom.

The A statistic provides a test of whether the J raters can be viewed as all the same (Alexander & Govern, 1994, p. 94). The A statistic is an omnibus test that indicates whether or not there are significant differences among the rater severities. Should the A statistic indicate significant differences among raters do exist, the value of the AG_{z_j} will be used to identify individual raters.

There exist a number of post hoc procedures that can be used to test for these differences. In this study, a procedure analogous to the running of multiple t-tests was employed. While this procedure has been criticized for excessive Type I error, this was not considered a flaw in the circumstances of this study. The base issue in any educational assessment is the fairness to students who have been assessed and about whom inferences are to be drawn. Thus, it is better to over identify raters who are either severe or lenient in relation to the other raters. In the case of severe raters, examinees are penalized if they receive too low a score. In contrast, in the case of lenient raters, examinees who must compete against other examinees who have been given erroneously high scores that are penalized. The use of a multiple t like procedure will help guard against these two potential misinterpretations.

The rater facet has been described as a random facet. If so, why attempt to isolate discrepant raters when in another replication of the assessment the same raters may not be involved? In the present context, it is true that the set of raters changes from one assessment to the next, although many raters do return for more than one marking session. It is also known that the marking of any one examination may take one to two weeks depending on subject area. This is more than ample time to identify and retrain raters whose severity error is considered too large to be tolerated.

Results

The A statistic value, 928.51, indicated the raters differed significantly from one another ($p < .0005$). Reported in Table 7 are the mean total scores by raters (rater severities), number of papers marked by the rater, rater standard error of the mean, and the Alexander Govern AG_{z_i} value for each of the 70 raters. A double line appears at the value of the AG_{z_i} statistic beyond which the probability of the value differing from zero by chance was less than or equal to .05, with shaded values at the .01 level of significance.

At the .05 level of significance, 37 (18 severe and 19 lenient) of the 70 raters (52.9%) were above (severe) or below (lenient) the cut off scores. As this is more than 10 times the expected number, 3 or 4, given the assumptions of equally severe raters and randomly selected papers, it is clear that, in spite of training, these raters varied in severity from their fellow raters. Likewise 29 (14 severe and 15 lenient) of the 70 raters (41.4%) fell beyond the cut offs at the .01 level, whereas only 1 rater might be expected to do so.

Table 7
 Rater Severities Ranked by Alexander-Govern z-Score

Rater Number	Rater Severity	Rater SD	Papers Marked	AG Standard error of Mean	AG z_i
58	22.88	4.59	121	0.417	7.534
57	24.33	4.56	212	0.313	6.419
5	24.27	4.03	150	0.329	6.180
59	24.17	5.37	222	0.360	6.046
36	23.73	5.28	154	0.425	5.999
68	24.37	5.57	275	0.336	5.972
10	25.12	5.39	244	0.345	3.780
52	24.51	5.89	138	0.501	3.754
1	25.11	4.35	136	0.373	3.492
65	25.15	4.91	178	0.368	3.455
21	25.12	5.17	173	0.393	3.313
32	25.45	4.19	172	0.319	3.068
50	25.07	6.50	183	0.480	2.827
31	25.29	6.20	210	0.428	2.674
30	25.52	4.72	145	0.392	2.334
35	25.77	4.65	258	0.289	2.318
70	25.54	5.68	177	0.427	2.104
69	25.75	4.97	211	0.342	2.019
56	25.72	6.91	302	0.398	1.817
20	26.01	5.43	397	0.273	1.593
54	26.02	4.81	252	0.303	1.399
33	25.84	5.67	149	0.465	1.297
16	25.91	6.67	192	0.481	1.109
23	26.13	5.27	233	0.345	0.911
18	26.02	6.32	178	0.474	0.895
14	26.01	6.16	116	0.572	0.758
37	26.04	5.59	109	0.535	0.754
17	26.12	5.99	186	0.439	0.739
67	26.24	4.80	291	0.281	0.728
49	26.22	5.57	285	0.330	0.682
19	26.19	5.50	161	0.438	0.581
41	26.24	5.63	188	0.411	0.499
48	26.32	5.70	437	0.273	0.459
38	26.25	5.54	164	0.433	0.450
3	26.24	5.69	143	0.476	0.430
43	26.30	5.35	331	0.385	0.377
9	26.41	3.96	284	0.235	0.149
47	26.42	6.56	84	0.715	0.035
4	26.73	5.39	291	0.316	-0.900
51	26.84	5.53	165	0.431	-0.915
66	26.90	5.78	142	0.485	-0.935

Rater Number	Rater Severity	Rater SD	Papers Marked	AG Standard error of Mean	AG $z_{.1}$
7	26.80	5.51	235	0.359	-0.985
11	26.78	5.81	299	0.336	-0.995
26	26.84	5.25	239	0.340	-1.160
8	26.84	5.10	237	0.331	-1.189
39	26.83	4.84	280	0.289	-1.327
6	27.00	5.37	207	0.373	-1.481
29	27.23	5.91	138	0.503	-1.550
63	27.09	5.63	229	0.372	-1.726
15	26.98	5.29	301	0.305	-1.748
55	27.14	6.39	276	0.385	-1.800
44	27.24	5.88	271	0.357	-2.213
13	27.48	6.41	213	0.439	-2.338
25	27.50	6.18	201	0.436	-2.399
24	27.76	7.07	172	0.539	-2.415
60	27.65	4.76	124	0.427	-2.769
45	27.27	5.09	305	0.291	-2.809
2	27.44	4.77	190	0.346	-2.840
28	27.48	5.26	215	0.359	-2.854
34	27.58	5.24	216	0.357	-3.143
62	27.88	5.55	154	0.448	-3.146
64	28.58	6.98	178	0.523	-3.984
27	29.31	7.13	111	0.677	-4.065
61	28.10	4.86	240	0.314	-5.125
53	28.67	5.76	201	0.406	-5.279
40	28.56	4.55	202	0.320	-6.277
12	28.82	4.46	177	0.335	-6.635
46	29.36	4.74	190	0.344	-7.793
42	29.76	5.78	277	0.347	-8.865
22	30.02	5.61	273	0.340	-9.634

These findings are similar to those reported by Braun (1988). Working at the .05 level of significance, he found 13 out of 36 raters (36.1%) who scored the English Literature and Composition Examination of the Advanced Placement Program (AP) had a large severity bias.

It is interesting to note that differences among adjacent raters in the ranked distribution are small and somewhat consistent. Application of the natural breaks

procedure (Sax, 1984) would suggest perhaps one break between the most severe rater and the next.

Detection of Central Tendency

Central tendency was assessed using a standard deviation measure, CTD, designed to detect dispersion of scale points awarded by a rater. Raters with little dispersion are said to possess central tendency.

Brown and Forsythe Test

To test for differences in dispersion, Ramsey (1994) suggested the use of a procedure proposed by Brown and Forsythe (1974) for leptokurtic distributions and a procedure developed by O'Brien (1981) for platykurtic distributions (p. 40). The Brown and Forsythe (BF) procedure was selected given that the total score distribution was leptokurtic (see Table 5). The BF procedure consists of replacing the mean used in Levine's test of homogeneity of variance with the median (Brown & Forsythe, 1974, p. 364) and then applying an ANOVA to these absolute deviations from the median.

Results

The results of the BF test are summarized in Table 8. As shown, the F statistic was significant, ($F_{69, 14720} = 5.85, p < .0005$), suggesting a lack of homogeneity of rater variances.

Table 8
Brown-Forsythe Summary ANOVA

Source of Variance	df	MS	F
Between Groups	69	73.8408	5.8497
Within Groups	14720	12.6231	
Total	14789		

However, as with the case for the Alexander-Govern test for rater severity, there are no specified testing procedures to follow up a significant BF test. To identify significantly different variances, the Chi-square test for a single variance with $n_j - 1$ degrees of freedom (Glass & Hopkins, 1984, p. 260) was used to construct a confidence interval around CTD_j^2 , the square of the central tendency deviation (see p. 38) on the mean CTD^2 . The mean CTD^2 was taken as the population value. CTD_j^2 s that differed significantly from this mean were considered to be a concern when scoring an examinee's paper. Table 9 contains the results of this test. The raters are ordered by their values of their CTD_j^2 s. The confidence limits for both the 95% and 99% intervals are listed, respectively, in the third and fourth columns. Confidence intervals that failed to span the mean CTD^2 are marked with an asterisk.

Of the 70 judges, 29 (41.4%) were identified as having a central tendency deviation at the .05 level of significance, in contrast to the 3 or 4 expected by chance alone. Of these 29, 16 raters were overly consistent in the use of marks close to their own mean mark. That is, these 16 raters displayed central tendency. The remaining 13 raters were significantly more willing to use a wide range of marks than the other raters. At the .01 level, 16 (22.9%) were identified as have significantly different CTD indices than the other raters; 10 displayed central tendency while 6 demonstrated a greater willingness to use the full range of marks.

Table 9
Confidence Intervals for Rater CTD

Rater	Rater CTD	Rater CTD ²	95% Confidence Interval		99% Confidence Interval	
5	5.786	33.482	5.272	6.413*	5.123	6.630*
9	5.892	34.711	5.283	6.661*	5.110	6.931*
32	5.970	35.640	5.531	6.485*	5.402	6.660*
1	6.289	39.547	5.648	7.094*	5.466	7.377*
40	6.298	39.666	5.773	6.928*	5.621	7.145*
57	6.319	39.934	5.771	6.984*	5.612	7.213*
54	6.328	40.042	5.711	7.096*	5.534	7.364*
58	6.393	40.868	5.901	6.975*	5.758	7.173*
30	6.411	41.096	5.822	7.133*	5.652	7.384*
12	6.515	42.442	5.891	7.287*	5.713	7.556
35	6.532	42.668	5.842	7.409*	5.646	7.719
24	6.536	42.719	6.052	7.104*	5.911	7.297*
46	6.571	43.182	5.964	7.317*	5.790	7.575
60	6.607	43.655	6.018	7.325*	5.848	7.573
67	6.719	45.141	6.222	7.302*	6.077	7.500
4	6.750	45.567	6.032	7.664	5.829	7.987
39	6.890	47.473	6.462	7.380*	6.335	7.544
2	6.982	48.748	6.451	7.609	6.296	7.822
26	7.002	49.028	6.506	7.580	6.361	7.776
61	7.005	49.068	6.320	7.858	6.124	8.156
8	7.085	50.190	6.290	8.110	6.067	8.475
6	7.088	50.238	6.471	7.835	6.293	8.093
62	7.109	50.537	6.383	8.022	6.176	8.344
15	7.137	50.940	6.601	7.770	6.444	7.984
59	7.152	51.152	6.487	7.971	6.296	8.255
43	7.191	51.718	6.550	7.973	6.366	8.243
19	7.236	52.353	6.640	7.950	6.467	8.194
36	7.363	54.212	6.498	8.495	6.256	8.901
29	7.400	54.765	6.618	8.394	6.396	8.745
41	7.402	54.792	6.853	8.048	6.692	8.268
49	7.420	55.057	6.780	8.194	6.595	8.461
45	7.421	55.073	6.774	8.206	6.587	8.476
20	7.436	55.287	6.825	8.168	6.647	8.419
21	7.441	55.374	6.826	8.179	6.648	8.432
52	7.447	55.455	6.466	8.781	6.195	9.270
37	7.450	55.499	6.746	8.319	6.544	8.621
70	7.472	55.825	6.601	8.608	6.357	9.016
7	7.488	56.069	6.886	8.206	6.711	8.451
48	7.493	56.141	6.914	8.177	6.746	8.411
23	7.509	56.383	6.754	8.456	6.538	8.788
11	7.534	56.768	6.936	8.247	6.761	8.491

Rater	Rater CTD	Rater CTD ²	95%		99%	
			Confidence Interval		Confidence Interval	
63	7.537	56.813	6.827	8.414	6.624	8.718
25	7.542	56.888	6.900	8.318	6.714	8.584
22	7.546	56.938	6.882	8.352	6.690	8.630
69	7.597	57.711	6.881	8.480	6.676	8.787
17	7.623	58.115	7.040	8.313	6.869	8.548
65	7.686	59.070	6.808	8.826	6.561	9.233
68	7.780	60.525	7.072	8.647	6.868	8.947
34	7.831	61.318	7.083	8.756	6.869	9.078
28	7.839	61.445	7.050	8.827	6.825	9.174
31	7.874	61.997	7.276	8.579	7.101	8.819
47	7.898	62.371	7.154	8.816	6.940	9.135
44	7.936	62.982	7.322	8.664	7.142	8.912
51	7.958	63.325	7.262	8.801	7.062	9.092
3	7.960	63.368	7.199	8.904	6.980	9.233
10	7.998	63.968	7.243	8.931	7.026	9.255
13	8.120	65.935	7.377	9.030	7.164	9.346
42	8.180	66.906	7.493*	9.006	7.294	9.290
50	8.202	67.274	7.493*	9.060	7.288	9.356
53	8.207	67.356	7.569*	8.963	7.383	9.221
38	8.251	68.074	7.566*	9.073	7.367	9.355
33	8.265	68.309	7.727*	8.884	7.568*	9.091
18	8.305	68.974	7.546*	9.236	7.327	9.559
66	8.406	70.661	7.758*	9.173	7.569*	9.434
14	8.421	70.909	7.487*	9.623	7.224	10.050
55	8.426	70.992	7.549*	9.534	7.300	9.925
56	8.583	73.663	7.883*	9.420	7.679*	9.707
16	8.632	74.512	7.983*	9.397	7.793*	9.657
64	8.816	77.718	7.896*	9.980	7.635*	10.390
27	9.610	92.356	8.663*	10.792	8.392*	11.206
Median	7.448		Mean	7.446		

Interrater Reliability

Correlations Among Raters

To determine the degree of interrater reliability, each rater's total scores for the papers marked were correlated with the total scores of the other raters who marked these papers. To obtain the mean of these rater correlations, the rater correlations were transformed to Fisher Z scores (Glass & Hopkins, 1984, p. 304). The mean Fisher Z

score was then computed, a 95% confidence interval constructed, and the lower and upper limits re-transformed to the corresponding correlation values. The correlation for the rater of interest reflected the degree of interrater reliability within raters who marked the same papers.

These rater intercorrelations and the respective confidence intervals are reported in Table 10. As shown, they ranged from .42 to .77, with a mean of .63. Values of the intercorrelation close to one indicate good interrater reliability within raters who marked the same papers; lesser values indicate a lack of reliability. As Nyberg (1987) described a rater as satisfactory if this interrater correlation was .8 or larger for English 33 data (p. 113), this standard was retained for purposes of comparability. A double line was added to Table 10 to indicate the raters that reach Nyberg's given sampling error. Only 6 of the 70 raters met this standard.

Interrater Reliability

Following the calculation of rater intercorrelation, the theoretical interrater reliability coefficient for the set of 70 raters where 3 raters are taken at a time was calculated. For calculation of this estimate of the reliability the Spearman-Brown prophecy formula was used:

$$\rho_{ij} = \frac{J\overline{\rho_{ij}}}{1 + (J-1)\overline{\rho_{ij}}},$$

where $\overline{\rho_{ij}}$ is the mean correlation across all raters, and

J the number of raters marking each paper

(Crocker & Algina, 1986, p. 119).

Table 10
 Intercorrelations Between a Rater and All Other Raters Who Marked a Common Bundle

Rater Number	Correlation	Lower Limit	Upper Limit
70	.77	.70	.82
59	.76	.70	.81
50	.76	.69	.81
32	.76	.69	.82
69	.75	.69	.80
63	.75	.69	.80
15	.74	.68	.78
56	.71	.65	.76
35	.71	.65	.77
2	.71	.63	.78
22	.70	.64	.76
23	.70	.63	.76
8	.70	.63	.76
19	.70	.61	.77
42	.69	.62	.75
39	.69	.62	.75
68	.69	.62	.75
16	.69	.60	.75
33	.69	.59	.76
29	.69	.59	.77
14	.69	.58	.77
27	.69	.58	.78
67	.68	.61	.74
61	.68	.61	.75
21	.68	.60	.76
31	.68	.59	.74
66	.68	.59	.76
37	.68	.57	.77
11	.67	.61	.73
55	.67	.60	.73
44	.67	.60	.73
49	.67	.60	.73
57	.67	.59	.74
7	.66	.58	.72
62	.66	.56	.74
10	.65	.57	.72
25	.65	.57	.73
41	.65	.56	.72
64	.65	.56	.73
47	.65	.51	.76
34	.64	.55	.71

Rater Number	Correlation	Lower Limit	Upper Limit
65	.64	.55	.72
20	.63	.57	.69
9	.62	.54	.69
28	.62	.53	.69
12	.62	.52	.70
36	.62	.51	.71
43	.61	.54	.67
13	.61	.52	.69
30	.61	.49	.70
1	.60	.48	.70
53	.59	.49	.67
24	.59	.48	.68
54	.58	.49	.66
40	.58	.48	.66
17	.58	.47	.67
51	.58	.46	.67
52	.58	.46	.68
60	.58	.45	.69
45	.57	.49	.64
38	.57	.46	.67
48	.56	.49	.62
26	.56	.47	.64
5	.56	.44	.66
6	.55	.45	.64
4	.53	.45	.61
18	.53	.42	.63
58	.49	.34	.61
46	.46	.35	.57
3	.42	.27	.54
Mean	.63		

Braun (1988) reported that in a study involving multiple raters the Spearman-Brown prophecy formula produced accurate estimates of essay grading reliability (p. 12).

One advantage of the interrater reliability approach is the ability to calculate a standard error of measurement, SEM or σ_e :

$$\sigma_e = \sigma_r \sqrt{1 - \rho_r} ,$$

where σ_e is the standard error of measurement,
 σ_x is the standard deviation of the distribution of scores, and
 $\rho_{jj'}$ is the interrater reliability for the j raters used in the scoring of the papers
 considered (Allen & Yen, 1979, p. 90).

The standard error of measurement can be used to construct a confidence interval which can then be used to estimate the range in which the true score is likely to be found.

Shown in Table 11 are the estimates of interrater reliability for the present study computed using the Spearman-Brown prophecy formula. In the first row is the mean correlation (see Table 10). Using this value the estimated reliability for three raters would be .84 (see Row 3, Table 11). The remaining results in Table 11 provide the estimated reliabilities for systems of 2, 4, and 5 raters. The right most column contains the SEMs associated with the interrater reliabilities. As expected, the reliability greatly increased with increased numbers of raters, and in a non linear fashion.

The Spearman-Brown prophecy formula was also used to place a lower limit for a rater's interrater correlation. If .80 is the acceptable standard that 3 raters as a group must attain, what is the minimum correlation that 3 raters of the same interrater reliability must possess? Application of the Spearman-Brown prophecy formula revealed a mean correlation value of .57 would be required. A double line was added to the lower portion of Table 10 to indicate which raters did not meet this standard when sampling error is ignored. This cut point indicated 9 raters did not meet this criterion. Shading was added to the correlation values in Table 10 to indicate which raters did not meet this standard even allowing for sampling error. This cut point indicated only 1 rater did not meet this criterion.

Table 11
Spearman-Brown Prediction of Reliabilities

Number of Raters	Reliability	SEM
1	.63	3.78
2	.78	2.95
3	.84	2.51
4	.87	2.22
5	.90	2.01

While the diminishing increases in reliability with increasing number of raters is clearly visible in Table 11, the corresponding values of the SEM are needed so that the impact of the changing number of raters can be judged. The use of only one rater would result in a 95% confidence interval that would be equal to 16.5% of the maximum score for the written examination, a clearly unacceptable margin of error. Even with the use of 3 raters, as was done in this study, the 95% confidence interval that would be equal to 10.9% of the maximum written examination score. The real worth of the SEM calculation, once the study has been carried out, is the ability to adjust cut scores to ensure that examinees are not harmed by miscategorization due to measurement error. For example, a pass or fail cut point could be adjusted to ensure that virtually no examinees were failed due to measurement error.

Coefficient alpha. Coefficient α does not take into account rater variation. Therefore, the use of coefficient α must result in a calculated reliability that is inflated when the data that are collected are based on raters making judgments. For the present study, coefficient α was calculated for each of the three separate markings. It was .90, with a SEM 2.01 in each case. The coefficient α was .89, SEM 2.08, when the median scores that Alberta Education employed were used to compute the final scores.

Comparisons with other studies. The mean interrater correlation, .63, is lower than the mean correlation of .73 reported by Becker, Hess, and Gibney (1993). However, the Spearman-Brown estimated reliability of .84 compared favourably with the reliability of .82 reported by Engelhard Jr. (1992). More importantly, the rater intercorrelations reported here were lower than the intercorrelations found by Nyberg (1987) in her study of the marking of the English 33 examination (see p. 29, Chapter III).

Correction of Rater Effects

Linear Scaling

As described in Chapter III, Hull's linear scaling was applied to the total score (see Equation 2, repeated here for clarity):

$$\hat{X}_{nj} = M_{CTD} \left(\frac{X_{nj} - M_{X.j}}{CTD_j} \right) + M_{...} \quad (2)$$

where $\hat{X}_{nj}$ is the adjusted total score for examinee n given by rater j,

M_{CTD} is the mean central tendency deviation of the group of raters,

X_{nj} refers to the score given by rater j to examinee n,

$M_{X.j}$ is the mean of rater j,

CTD_j is the central tendency deviation of rater j, and

$M_{...}$ refers to the mean score given by all raters.

With this scaling the examinee's observed score would have 3 corrected scores given each was marked by three raters. To get the corrected score for an examinee, the three corrected scores were averaged to produce a corrected total score for examinees. This corrected score was then rounded to the nearest unit and compared to the uncorrected total score.

Table 12 gives the percentage of differences between uncorrected and corrected total scores. As shown, the differences ranged from -4 to +3 points with 56.6% of the pairs of scores differing by one or more points; 10.6% by two or more points, and 1.2% by more than 2 points.

Table 12
Frequency of Score Differences

Score Difference	Percent
-3.00	0.2%
-2.00	3.4%
-1.00	22.4%
0.00	45.8%
1.00	24.7%
2.00	3.3%
3.00	0.0%

The distribution of the score differences was summarized by two statistics. The first was the conditional root mean square, CRMS,

$$CRMS = \sqrt{\frac{\sum_{n=1}^N (X_c - X_u)^2}{N - 1}},$$

where CRMS is the standard deviation of the differences between the corrected and uncorrected scores,

X_c is the corrected mean total for examinee n ,

X_u is the uncorrected mean total for examinee n , and

N is the total number of examinees.

The second statistic used was the mean absolute difference between corrected and uncorrected totals, also known as the average absolute difference or AAD:

$$AAD = \frac{\sum_{n=1}^N |X_c - X_u|}{N}$$

where AAD is the average absolute deviation of the differences between the corrected and uncorrected scores,

X_c is the corrected mean total for examinee n ,

X_u is the uncorrected mean total for examinee n , and

N is the total number of examinees.

The value for the CRMS for the linear scaling was 0.82; the AAD was 0.65.

Both the CRMS and the AAD corresponded with what is shown in Table 12; most of the corrected scores did not vary by more than one point from the uncorrected score and virtually all the corrected scores were within two points of the uncorrected score.

The correlation between the corrected scores and the uncorrected scores was .99. The uncorrected score and the corrected score both correlated .56 with the multiple choice portion of the English 33 examination.

While the value of the CRMS and the AAD suggest that overall there appears to be close agreement between the uncorrected and corrected scores, the fact is that over one half the examinees had at least a one point difference. As described in Chapter I, a one point error due to a single flawed multiple choice item is viewed seriously by Alberta Education; the item is omitted. A score difference that is unacceptable for one portion of the examination cannot be acceptable for another portion of the examination. The *Principles for Fair Student Assessment Practices for Education in Canada* (1993) state that all examinees must be treated fairly and equitably (p. 3). Therefore, it is recommended, in agreement with Nyberg (1987), that a mathematical correction be employed to reduce the

inequity based on a mean score that occurs whenever examinees are rated by a few raters from a large pool of raters.

CHAPTER V GENERALIZABILITY THEORY APPROACH

The procedures and results for the generalizability theory approach are presented in this chapter. First, two initial considerations concerning, (a) the level of analysis and (b) the nature of the data matrix are discussed. Second, a preliminary analysis and the corresponding results are described. The preliminary analysis was carried out to establish the feasibility of the proposed method of obtaining variance estimates by the aggregation of the estimates obtained by the analysis of individual bundles of papers. Third, a description of the main analyses and their results is presented. Included in the main analyses are the aggregation of bundle results to obtain stable and precise variance component estimates and the interpretation of these components, the illustration of the identification of causes of large variance components for the interactions, and the determination of generalizability coefficients and dependability coefficients. Lastly, several D studies, were carried out to predict outcomes of possible changes to the system.

Computer Programs Employed

The generalizability analyses were conducted using the BMDP program 8V available through BMDP Statistical Software Inc. and documented in the *BMDP Statistical Software Manual Volume 2* (BMDP Statistical Software Inc., 1990). The sampling procedures were carried out using programs from SPSS-X available through the SPSS-X Data Analysis System, Release 4.0, as documented in the *SPSS Reference Guide* (SPSS Inc., 1990).

Initial Considerations

Two issues related to the nature of the test and marking design needed to be considered before proceeding with the generalizability analysis. The first issue concerned the consideration of scale, the lowest level of marking, as a facet in the design rather than the section. The second issue concerned the relatively empty data matrix.

Scales

Scale, consisting of nine levels corresponding to the 9 scoring scales used when marking papers, was included as a facet in the present analysis. As revealed in Table 5, the intercorrelations among the scales were moderate to moderately high in value, suggesting that although the scales are related, one scale was not a simple transformation of another, that is, the scales were distinctive.

Section was not considered a facet. By ignoring this facet a balanced design was realized. Given the difficulties encountered in completing a G analysis for an unbalanced design (Elder, 1991), use of a balanced design was preferred. Further, if a large variance component attributable to scale or to the interaction of scale with the other facets were found, the corresponding means could be inspected to see if they were related to section. Consequently, section was dropped as a facet.

Relatively Empty Data Matrix

The second issue concerns the data matrix. As described in Chapter II, the data matrix is very sparsely filled, 4.3% full. It can be described as consisting of a series of $n \times j \times i$ (examinee-by-raters-by-scales) submatrices. While the full data matrix is not amenable to generalizability analysis, these submatrices are. Consequently, the decision was taken to conduct the analyses at the submatrix level (see Chapter III).

Preliminary Analyses and Results

The data analysis plan for the generalizability analysis called for a set of replicated analyses of a sample of $n \times j \times i$ submatrices. Given n was 6, the estimated variance components would likely be somewhat unstable. However a question arose, would the means of the variance components be sufficiently stable to warrant their interpretation and subsequent use? Consequently a preliminary analysis was first conducted to assess the reasonableness of the mean estimates obtained. Detailed analysis and results of these analyses are presented in Appendix E. A summary of these analyses and results is presented here.

Sample. The sample was selected using the third marker identification number. Since rater identification numbers were considered likely to produce a random sample of the examinee population while examinee identification numbers or examinee record numbers would group examinees according to school and region, rater number was considered an appropriate variable for selecting a sample. Given that the number of possible marker triplets far exceeded the number of bundles, the rater number was the one identification number that would result in the selection of combinations of three raters that actually did mark a common bundle of papers. The third rater was arbitrarily picked as the selection variable.

A sample was drawn so that 100 intact bundles of six were chosen. This sample contained all intact bundles for raters 1 to 12 plus sufficient bundles from rater 13 to form the 100 bundles for this sample. This sample was divided into 5 units of 20 intact bundles of six to provide 5 "replicated" studies. As examinees were not matched with raters in any known way as bundles were selected haphazardly by raters, this preliminary sample, comprising 12.2% of the examinee population, should be a representative sample of examinee proficiencies, rater severities, and scale difficulties.

Methods and Results. Each bundle within each replicated study was analyzed according to a fully crossed, $n \times j \times i$, examinees-by-raters-by-scales random effects design. The percentage of total variance was calculated for each component within each bundle. Then within each replication, the mean percentages across the 20 bundles were computed

It was found that variance estimates of individual bundles within a replication varied widely as expected. However, the mean variance estimates, while different across replications, maintained essentially the same ranking by magnitude across the five replications. It was concluded that this method of obtaining population estimates by aggregating estimates produced by analysis of bundles was a feasible method for analyzing the data matrix.

Additional preliminary analysis. To examine whether the results would be different for broken bundles, and bundles that have only two common raters, the analyses were repeated using these bundles. The results were similar to the use of the bundles of six. Consequently, only bundles of six were employed for the main analyses. More detailed results and the conclusions of this preliminary analysis are contained in Appendix E.

Main Analyses and Results

The Sample

The bundle selection process for the main sample was a continuation of the process employed for the preliminary sample. A sample of 177 intact bundles was used for the main analysis. This number was required to ensure that no rater was sampled less than three times, although some raters were more heavily sampled than others. The corresponding number of examinees, 1062, comprised 21.6% of the examinee population. As discussed in the description of the preliminary study this number of

examinees, not selected according to any known characteristic of the examinees, was expected to be representative of the population of examinees.

As shown in Table 13 the sample of bundles was selected in groups, with the first eight groups containing 20 bundles; the ninth group contained 17. The total number of observations, 28, 764, ($1062 \times 9 \times 3$) exceeded the minimum suggested by Smith (1978) to provide variance estimates that would be reasonably precise estimates of the corresponding population components being studied.

Table 13
The Generalizability Group Composition Characteristics

Group Number	Bundle			
	Number	Size	Quantity	Total N
1	1-20	6	20	120
2	21-40	6	20	240
3	41-60	6	20	360
4	61-80	6	20	480
5	81-100	6	20	600
6	101-120	6	20	720
7	121-140	6	20	840
8	141-160	6	20	960
9	161-177	6	17	1062

Estimation of Variance Components

Method and Results. Each bundle was analyzed according to a fully crossed $n \times j \times i$ design. The mean variance estimate for each facet was computed for each group and for the total (177 bundles). The variance component estimates for each group and for the total, including the percentage variance estimates for the total, are given in Table 14. Presented in Table 14 are the variance components for: examinees (N), raters (J), scales (I), examinee-by-rater interaction (NJ), examinee-by-scale interaction (NI), rater-by-scale interaction (JI), and the examinee-by-rater-by-scale interaction confounded with error (NJI,E).

Table 14
Variance Components for the Groups and Total Sample

Group	N	J	I	NJ	NI	JI	NJI,E
1	0.159	0.043	0.008	0.096	0.190	0.028	0.253
2	0.225	0.050	0.011	0.074	0.099	0.020	0.210
3	0.178	0.032	0.008	0.057	0.146	0.025	0.203
4	0.262	0.034	0.015	0.109	0.124	0.020	0.232
5	0.208	0.021	-0.002	0.085	0.135	0.031	0.226
6	0.217	0.013	0.000	0.065	0.148	0.024	0.213
7	0.358	0.027	0.002	0.078	0.119	0.032	0.229
8	0.283	0.086	0.018	0.077	0.152	0.014	0.238
9	0.214	0.078	-0.003	0.091	0.160	0.023	0.237
Total 1-9	0.234	0.035	0.006	0.083	0.127	0.024	0.226
Total %	31.84%	4.76%	0.82%	11.29%	17.28%	3.26%	30.75%

While variation occurred among the group mean estimates produced within the groups, the ranking of variances by size was essentially consistent across groups. This result is displayed in graphical form in Figure 3. As shown, the ordering by size of the variance estimates for the different facets is relatively consistent across groups. Consequently, the variance component estimates for the total were retained and used in the analyses that follow.

As shown in the last row of Table 14, the examinee component and the examinee-by-rater-by-scale error component accounted for the largest percentages of variance, respectively accounting for approximately 31.8% and 30.8% of the total variance. The next largest components were the examinee-by-rater and examinee-by-scale interactions. These components accounted for 11.3% and 17.3% respectively. The remaining components, rater, scale, and the rater-by-scale interaction were low, accounting for 4.8%, 0.8%, and 3.3% of the total variance respectively.

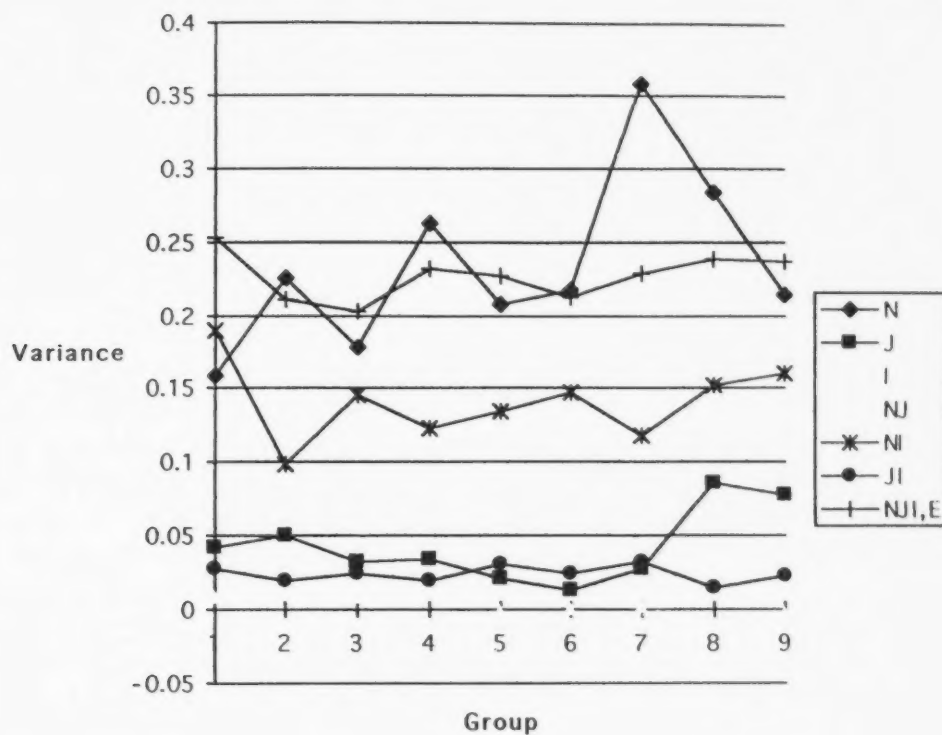


Figure 3.
Variance estimates for components by group.

Interpretation of the Magnitude of Variance Components

The examinee (N) facet variance component, σ^2_n , reflects differences among examinees. This component accounted for a large proportion of the total systematic variance considered in this study. This is a desirable result as the examination was designed to detect examinee differences. The rater facet (J) values, σ^2_j , while much smaller than the examinee facet values, indicated that some bundles contained a rater who differed greatly from his or her fellow raters; there are rater severity differences. The scale facet (I) values, σ^2_i , showed very little difference among the scales. In contrast, the interactions involving examinees with each of the other two facets were relatively greater than the rater and scale facets and therefore troublesome. The relatively large

examinee-rater (NJ) interaction variance component, σ^2_{nj} , suggested raters of the same examinees ranked these examinees differently. This may have been an indication of halo effect as defined by Gronlund and Linn (1990) and Popham (1990) (see also page 4) in that two raters may not have seen differences within an examinee's performance that the other rater of the examinee had seen. The examinee-scale (NI) interaction variance component, σ^2_{ni} , suggested that examinee performance was not consistent across the nine scales. The rater-scale (JI) interaction variance component, σ^2_{ji} , while somewhat smaller than the previous two interaction components, suggested that some raters behaved quite differently across the scales. Lastly, the large examinee-rater-scale error term, $\sigma^2_{nji,e}$, which included both random and systematic sources of variation, may have been due to a complex interaction among raters' views of examinees' responses on certain scales or to an important source of variation that was not included in the design used in this study.

Identification of Raters Who Differ from Other Raters

While classical test theory allows only the consideration of each facet separately, generalizability theory allows the analysis of interaction effects (Crowley, Thompson, and Worchel, 1994, p. 706). In this study, while the variance due to raters accounted for a relatively small percentage (4.8%) of the total variance, and the variance due to scales accounted for an even smaller percentage (0.8%), the variance due to the interaction between examinees and scales and the variance due to the interaction between examinees and raters both accounted for relatively larger percentages (17.3% and 11.3%, respectively).

One bundle, consisting of a triplet of raters, was chosen to illustrate the information that can be gained from the examination of these interactions. The variance components for the object of measurement and for all facets are presented in Table 15. For this particular triplet, the rater, scale, and the rater-by-scale interaction variance components were all

small, while the examinee-by-rater interaction and examinee-by-scale interaction components were both relatively large. These results are similar to the total sample results in Table 14.

Table 15
Variance Components for Triplet 63

N	J	I	NJ	NI	JI	NJI, E
0.355	0.005	0.001	0.185	0.083	0.028	0.131

The deviation scores for the members of each facet were examined. Deviation scores were used as these scores more clearly illustrate facet and interaction differences than do raw cell scores. The deviations for the examinee scores, rater scores, and scale scores are given in Table 16 while the deviation scores for the interactions are found in Tables 17, 18, and 19. In these tables the examinees are labeled 1 to 6 while actual rater numbers and scale names are retained.

As shown in Table 16, the examinee deviations ranged from 0.796 to -0.685, consistent with the large variance for examinees. The scores awarded to examinees 1, 2, and 6 were higher than the scores awarded to examinees 3, 4, and 5. It is assumed that the differences in scores reflected the actual differences in writing achievement of these six examinees. The rater deviations ranged from 0.129 to -0.204 consistent with the small variance for raters. Rater 9, the third rater, awarded slightly lower scores than did the other two raters. The scale deviations ranged from 0.241 to -0.370; all but two scales were located within the range of 0.074 to -0.092. The third scale, persMCO, was given somewhat higher scores while the fifth scale, funcTD, had somewhat lower scores than the other scales. With the exception of the third and fifth scales, the differences among scales were all small, consistent with the small variance for scales found in this triplet and found in the G study as a whole.

Table 16
Deviation Scores for Examinees, Raters, Scales

Examinee								
1	2	3	4	5	6			
0.648	0.796	-0.611	-0.426	-0.685	0.277			
Rater								
2	7	9						
0.129	0.074	-0.204						
Scale								
1 pers TD	2 pers ORG	3 pers MCO	4 pers MCH	5 func TD	6 func WRS	7 vis TD	8 vis ORG	9 vis WRS
0.074	0.074	0.241	0.129	-0.370	-0.037	-0.037	-0.092	0.019

As shown in Table 17, the source of the large examinee-by-rater interaction was apparent when the deviations were examined. Deviations ranged from 0.500 to -0.667. The first two raters were more inconsistent than the third. The first rater, rater 2, rated two examinees (1 and 6) relatively high while three examinees (3, 4, and 5) were rated relatively low. The second rater, rater 7, rated the fourth examinee high and the sixth examinee very low in direct contrast to the first rater. The third rater, rater 9, showed more consistency than the other two raters.

Table 17
Deviation Scores for Examinee by Rater Interactions

Examinee by Rater Deviations			
Examinee	Rater		
	2	7	9
1	0.463	-0.250	-0.204
2	0.093	0.148	-0.241
3	-0.389	0.000	0.389
4	-0.352	0.481	-0.130
5	-0.315	0.296	0.019
6	0.500	-0.667	0.167

Differences in an examinee's behaviour on individual scales are apparent. Examination of Table 18 reveals that the examinees exhibited the most variability on the fifth scale, funcTD, and the seventh scale, visTD.

Table 18
Deviation Scores for Examinees by Scale Interaction

Exam inee	Examinee by Scale Deviations								
	1 pers TD	2 pers ORG	3 pers MCO	4 pers MCH	5 func TD	6 func WRS	7 vis TD	8 vis ORG	9 vis WRS
1	0.074	-0.259	-0.092	0.018	-0.148	0.185	-0.148	0.241	0.130
2	0.259	0.259	0.426	0.204	-0.963	-0.300	0.370	-0.241	-0.019
3	0.000	-0.333	-0.167	-0.056	0.111	0.111	0.111	0.167	0.056
4	-0.518	0.148	-0.019	0.093	-0.407	-0.047	0.593	-0.093	0.129
5	0.074	-0.259	-0.093	0.018	0.519	0.185	-0.481	-0.093	0.130
6	0.111	0.444	-0.056	-0.278	0.888	-0.111	-0.444	-0.056	-0.500

For example, examinee 2 had an unexpectedly low score for the fifth scale, funcTD, while the sixth examinee had an unexpectedly high score. Other examinee by scale differences are apparent in Table 18 but as examinees are expected to display some variation in the traits they exhibit, this feature was not considered a flaw of the rating system.

The rater-by-scale variance was small suggesting as illustrated in Table 19 that the three raters displayed relatively little variation across scales (columns). For example, with the exception of two scales for rater 7 (scales 5 and 6) and three scales for rater 9 (scales 6, 7, and 8), the raters were all consistent across scales.

Table 19
Deviation Scores for Rater by Scale Interaction

Rater by Scale Deviations									
Rater	1 pers TD	2 pers ORG	3 pers MCO	4 pers MCH	5 func TD	6 func WRS	7 vis TD	8 vis ORG	9 vis WRS
2	0.092	-0.241	0.093	-0.130	-0.130	-0.130	0.203	0.093	0.149
7	-0.018	0.148	-0.185	0.093	0.259	-0.241	0.093	0.148	-0.130
9	0.093	0.093	0.093	0.037	-0.130	0.370	-0.296	-0.241	-0.018

Interaction Effects for the Entire Sample

The results of one bundle were used to illustrate the differences among members within a facet. Given the development of appropriate software, larger amounts of time, or a much smaller sample, it would be possible to compile the deviations for each rater across bundles. However, such an analysis with a relatively empty matrix such as those found in large scale testing programs like the one considered in this study will likely not be feasible.

Decision Studies

In addition to describing the sources of variability in a testing situation, the variance components can be used in a decision study, or D study (Cronbach et al, 1972). These studies involve the calculation of a variety of reliability like coefficients and their corresponding standard errors of measurement which take into account the nature of the design and the generalization to be made. Often the variance components are used to plan the decision study which is conducted later. Other times, in the interest of saving time and money, the results from the G study are directly used in the D study. Such is the case with Alberta Education. Consequently, the D statistics for the data set are presented first. Then the results of the G study are used to plan and assess alternative D study designs.

Use of G Study Results Directly in a D Study

Both coefficients for a relative decision – generalizability coefficients, and coefficients for absolute decisions – dependability coefficients, were calculated. While Alberta Education computes only the first, decisions are made that correspond more to the second as well. Two values for each coefficient were computed. The first corresponds to the procedure followed by Alberta Education. Equivalent to Cronbach's α for a single test, it is computed using analysis of variance for an $n \times i$ design (Hoyt, 1941). The second coefficient reflects more properly the actual testing design, incorporating the rater facet and taking into account variability among bundles or sets of papers marked by the same raters (see Table 2).

The results of the calculations are reported in Table 20 for both the generalizability and dependability coefficients together with their corresponding standard errors of measurement, δ and Δ . Linn and Burton (1994) recently reminded test developers that it is misguided to focus on only generalizability and dependability coefficients and indicated that

the corresponding standard errors of measurement (SEM) should be reported. In Table 20, the results determined using Hoyt's procedure are labelled Hoyt's ANOVA; the results determined incorporating the rater facet and variation among bundles are referred to as the Full analysis results. In the case of the dependability coefficients, a cut off score of 50% was used. This score corresponds to the passing standard set by Alberta Education for the total test (the total of the multiple choice and essay components).

Table 20
Values of Generalizability and Dependability Coefficients

	Generalizability (relative decision)		Dependability coefficient (absolute decision)	
	Coefficient	SEM (δ)	Coefficient	SEM (Δ)
Hoyt's ANOVA	.9031	1.76	.9008	1.78
Full analysis	.7911	2.58	.7873	2.60

Examination of Table 20 reveals that, for a relative decision the value of Hoyt's ANOVA is .90 with a corresponding SEM of 1.76. When variation due to raters and bundles was taken into account the generalizability coefficient dropped to .79 with a corresponding increase in SEM to 2.58. The corresponding dependability coefficients are slightly lower than their generalizability counterparts due to the presence of an additional component in the error variance and the observation that the observed mean score does not equal the cut off score used.

When computing the values for the dependability coefficient, a cut off score of 50% was employed; this corresponded to the passing score used by Alberta Education. However, this agency fails to incorporate this cut off score when it computes its estimate of reliability and the standard error of measurement estimates. These estimates, as pointed out in the previous chapter, are coefficient α and its related standard error of

measurement, SEM. These values are .90 and 1.76, respectively, as reported in Table 20. However the correct values are respectively .79 and 2.60.

Impact of differing estimates of SEM. To see the consequences of the difference, consider the lower limit of the 95% confidence interval built around the cut off score and below which students are failed (those above the lower limit are considered to be below the cut off score due to measurement error). Using the present SEM, students who are below 46% are failed. Using the more appropriate SEM, students who score below 44.8% or below fail. Clearly the number in the latter case will be less than the number failed in the former case.

Use of G Study Results to Plan and Assess Alternative D Studies

The estimated variance components for the facets selected in a generalizability study (G study) can be used to design a measurement for a particular purpose in a D study. Generalizability theory allows for the possibility of calculation of coefficient values and related SEMs for levels of facets in a D study that differ from the number of the levels in the original G study. The process is analogous to the use of the Spearman-Brown procedure for estimating reliabilities for varying numbers of examination items or varying numbers of raters in a classical test score analysis.

A series of decision study designs was examined to investigate the effects of varying the numbers of raters and the number of scales. The number of raters studied was varied from 1 to 5, and the number of scales from 2 to 9. Given the number of possible combinations, 45, to be analyzed is large, selected combinations were analyzed. The rationale for each choice is given with the analysis for that choice. The analyses for varying the numbers of raters is presented first, followed by the analyses for varying the numbers of scales. The section concludes with an example in which the effect of

reduction of the number of raters is compared with the effect of the reduction by the comparable number of scales.

Varying the Number of Raters

In the first set of analyses the number of raters was varied from one to five while the number of scales was kept at nine. The number of raters was varied from one to five as it is conceivable that the examining agency would wish to lower the number of raters to reduce cost or be forced to increase the number of raters due to low interrater reliability. An upper bound of five was thought to reflect the maximum cost an examining agency would be willing to incur. The generalizability coefficients and dependability indices are located in Table 21. Again coefficients are provided for both the Hoyt's ANOVA and Full analysis interpretations. An asterisk, *, is placed in the column for 3 raters as this was the number of raters used in the study considered in this research.

Table 21
D Study – Various Numbers of Raters

	Number of Raters				
	1	2	3 *	4	5
Generalizability Coefficients					
Hoyt's ANOVA	.9031 (1.76)	.9031 (1.76)	.9031 (1.76)	.9031 (1.76)	.9031 (1.76)
Full analysis	.5982 (3.58)	.7301 (2.93)	.7911 (2.58)	.8243 (2.36)	.8456 (2.22)
Indices of Dependability					
Hoyt's ANOVA	.9008 (1.78)	.9008 (1.78)	.9008 (1.78)	.9008 (1.78)	.9008 (1.78)
Full analysis	.5931 (3.98)	.7275 (2.94)	.7873 (2.60)	.8204 (2.39)	.8420 (2.24)

As the Hoyt's ANOVA calculation disregards rater variation, there were no changes in values for both the generalizability coefficient and the dependability coefficients, with a change in the number of raters. This shortcoming clearly indicates the inadequacy of this calculation for a system that includes raters producing scores on scales. In contrast, there were changes in the Full analysis in which variation due to raters and bundles was explicitly accounted for. However, and not unexpectedly, the amount of change decreased with increasing number of raters. The changes in values of the full analysis coefficients were greatest when the number of raters increased from 1 to 2 raters, .10 or greater; and were less when the number of raters increased from 4 to 5.

The trend for the SEMs is the same as for the generalizability coefficients, although in the reverse direction. That is, the change in values of the SEMs was greatest when the number of raters increased from 1 to 2 raters, 1.0 to 1.5 score points; and was smallest when the number of raters increased from 4 to 5.

Varying the Number of Scales

In the second study the number of raters were kept at three while scales were varied from two to seven. As the current examination of nine scales had a sitting time of two and one half hours, an increase in the number of scales due to another writing task being added would result in an increase in writing time. The necessary increase likely would not be considered by the testing agency. In predicting coefficient values resulting from the reduction of scales, the assumption must be made that the scales are all similar. These results are presented in Table 22.

Table 22
D Study – Various Numbers of Scales

	Number of Scales					
	2	3	4	5	6	7
Generalizability Coefficients						
Hoyt's ANOVA	.6744 (3.22)	.7565 (2.78)	.8055 (2.49)	.8351 (2.29)	.8614 (2.10)	.8788 (1.96)
Full analysis	.6249 (3.45)	.6867 (3.16)	.7224 (2.97)	.7457 (2.84)	.7621 (2.75)	.7743 (2.68)
Indices of Dependability						
Hoyt's ANOVA	.6686 (3.25)	.7516 (2.81)	.8014 (2.51)	.8345 (2.29)	.8582 (2.12)	.8745 (2.00)
Full analysis	.6134 (3.51)	.6774 (3.20)	.7147 (3.01)	.7391 (2.88)	.7563 (2.78)	.7691 (2.71)

First it should be noted that unlike the case for raters, scales are explicitly recognized in the calculation of Hoyt's ANOVA. Consequently both the generalizability coefficient and the dependability coefficient values will change with a change in the number of scales. Changes in values of these coefficients were greatest from 2 to 3 scales, generally .06 or greater; changes are less for 3 to 4 scales, and changes are relatively unimportant for more than 4 scales. The trends described for the SEMs are the same as previously described for the coefficients. Increasing the number of scales from 2 to 3 decreases the width of the confidence interval by 0.6 or more score points; while increasing the number of raters from 3 to 4 decreases the width of the confidence interval by a lesser degree, and increasing the number of scales to 5 or more results in little further decrease in SEM.

An Application of the D Study

In a decision study any number of conditions for each facet could be considered simultaneously. Most of the conditions chosen here were selected for the purpose of

examining cost cutting measures. To illustrate, the question asked could be: "If marking time, and hence costs, are to be reduced by one-third, will higher reliability be obtained by the use of only two raters, or by the shortening of the examination to two sections of six scales total?"

The value of the full analysis dependability index for two raters on nine scales is .73, SEM 2.94; while for three raters on six scales the value of the full analysis dependability index is .76, SEM 2.78. These two examples illustrate that, for the data in this data set, reducing the length of the task reduces reliability less than reducing the number of raters that mark a paper. As the reduction of the number of raters by one-third produced the same level of error as reducing the number of scales to less than one-quarter (two-ninths) the original value, a reduction in the length of the examination is likely to be the more cost efficient measure.⁵ It should be noted though, the D study cannot address the question of whether the reduction of scales would result in a reduction in the match of examination content and course curriculum.

⁵ As described in Chapter II, p. 20, Alberta Education has reduced the number of markings to 2 per paper; the number of scales has been reduced to 8. However, this reduction in scales was not done by a reduction in number of writing tasks but by a dropping of one scale.

CHAPTER VI MULTIFACETED RASCH MODEL RESULTS

The procedures and results for the multifaceted Rasch model approach (Linacre, 1989), the third approach considered in this study, are presented in this chapter. First, a series of preliminary considerations and analyses were examined prior to the main Rasch analyses being carried out. These preliminary analyses included an examination of dimensionality, discrimination of scales, and speededness. A misfit analysis was then completed just prior to the main analysis. The misfit analysis included an examination of misfit by scale and misfit by scale point.

The misfit analysis and main analyses employed a three facet model in which the facets were examinee proficiency, rater severity, and scale difficulty. The focus of the study was on rater behaviour. As such, the majority of the discussion of the analysis is centered on the rater statistics. However, as raters rate examinees and as raters' responses to examinees' writing determine scale characteristics, some discussion of examinee characteristics and scale characteristics takes place. Three sets of statistics were produced for individual members of each facet: a logit measure of the facet characteristic (e.g., rater severity), infit and outfit statistics, and Rasch point biserials. A series of indices that describe the facet as a whole, for example, reliability of separation index, were also employed and are presented and discussed in the appropriate subsection. The description of the facets is followed by a comparison of the Rasch fair average, which is an examinee score corrected for rater and scale, with the observed mean score for the examinee.

Computer Programs Employed

Two Rasch programs were employed in this study. The data required reformatting to a file structure suitable for analysis by the multifaceted Rasch analysis computer program. The program used for reformatting was Facform, Version 1.22, described in *A User's Guide to Facform* (Linacre, 1992). The computer program used to carry out the multifaceted Rasch analyses was Facets⁶, version 2.6, (Linacre, 1992). Both programs are described in *FACETS: Many-Facet Rasch Analysis* (Linacre, 1992). The factor analysis was carried out using the Factor program within the SPSS-X Data Analysis System, Release 4.0, as documented in the *SPSS Reference Guide* (SPSS Inc., 1990). In addition, Microsoft Excel 4.0, as documented in *Microsoft Excel User's Guide 1* (Microsoft Corporation, 1992), *Microsoft Excel User's Guide 2* (Microsoft Corporation, 1992), and *Microsoft Excel Function Reference* (Microsoft Corporation, 1992) were employed.

Preliminary Analyses and Results

Suitability of the Data Matrix

First, examination of the data matrix revealed that, while the data matrix was sparsely filled (see Chapter II), there was sufficient linkage within the data matrix to permit the estimation of the Rasch parameters associated with the examinee, rater, and scale facets in the design. There was no self contained sub set of raters that marked only a set of papers that other raters outside the sub set did not encounter.

⁶ This program does not allow the user to alter the number of significant digits for any of the numbers that are reported as part of the Facets output. For example, Observed average and Fair average are reported to 1 decimal place while logit values and fit statistics are reported to 2 decimal places, and t statistics are reported as integers.

Assumptions

Next, the assumptions underlying the use of the Rasch model were assessed. The assumptions that were examined were unidimensionality, equal discrimination, and nonspeededness. As the responses were in supply format rather than selection format, the assumption of guessing was not considered.

Dimensionality

In the present context, dimensionality refers to the factorial composition of the total score. The multifaceted Rasch analysis is based on the notion that this composition is essentially unidimensional. As evidence of essential unidimensionality, Nandakumar (1994) suggested that the presence of only one dominant factor or dimension satisfied the assumption (p. 18). While historically linear factor analysis has been used to assess dimensionality, Nandakumar (1994) cautioned that "there are a number of technical and methodological problems ... item difficulty and guessing" (p. 18). As difficulties for the nine scales were approximately equal, ranging from 2.775 to 3.076 (see Table 4), scale difficulty was not considered an issue and, as just previously described, the assumption of guessing was not considered as the examination consisted of writing tasks. Therefore a linear factor analysis was employed to test for essential unidimensionality.

A principal components factor analysis was conducted. The eigenvalues of the components, along with percent variance accounted for, are presented in Table 23. As shown in this table, the first factor accounted for 54.7% of the total variance, while the second accounted for 14.0%. The first factor is nearly four times larger than the second. Reckase (1979) concluded that "for successful calibration, the first factor should account for at least 20% of the test variance" (p. 228), while much later, Huynh and Ferrara (1994) later suggested "good ability estimates can be obtained even if the first component accounts for less than 10% of the total test variance" (p. 127). The first factor in the

present study is clearly dominant and accounts for over half the total variance. Thus it was concluded that the data can be considered essentially unidimensional.

The Assumption of Equal Discrimination

Hambleton and Murray (1983) suggested that the identification of items that have scale-test score correlations that are within some specified range, e.g., .15, be considered evidence that the discrimination indices are equal (p. 75). In the present study, the mean corrected correlation was .66 with a standard deviation of .06 (see Table 5). With the exception of the fifth scale, which correlated somewhat less than the others, all others had scale-test score correlations within of .08 of the mean scale-test score correlation. The scales were considered to have equal discrimination.

Table 23
Eigenvalues for all Nine Factors

Factor	Eigenvalue	Percent Variation	Cumulative Percent
1	4.92096	54.7	54.7
2	1.25857	14.0	68.7
3	0.84505	9.4	78.1
4	0.68829	7.6	85.7
5	0.30794	3.4	89.1
6	0.27570	3.1	92.2
7	0.25095	2.8	95.0
8	0.23788	2.6	97.6
9	0.21468	2.4	100.0

The Assumption of Nonspeededness

Examinees who did not respond to a writing task were awarded a score of 0 on each scale used to mark the responses to that task. Examination of the marks awarded to examinees revealed that the percentage of 0s awarded by scale ranged from a 0.3% to 1.5%. The very low rate of non response indicated that all but a small number of examinees completed every writing task and therefore speededness was not a factor.

Data Misfit Analyses and Results

As done in all Rasch analysis to solve problems of indeterminacy in estimates, examinees who were awarded either perfect scores or zero scores on all nine scales by all of the three judges who marked them were excluded from the analyses. Three such examinees were found, resulting in a calibration sample of 4,927 examinees.

Next, a misfit analysis, employing the infit statistic, was carried out to assess the fit of the data to the model. Given that there were 4,927 examinees rated on 9 scales by 3 raters, there were 133,029 responses judged for the degree of misfit. The infit statistic, judged by use of a standardized residual t-statistic of value greater than or equal to the absolute value of 3, was used as the criterion for the selection of misfit responses. Use of this criterion designates approximately 1.0% of the responses as misfit if the data fit the Rasch model (Linacre & Wright, 1992, p. 42). Of the 133,029 responses, 0.7% were found to be misfitting. Portions of the misfit response table are reproduced in Table 24.

Following the advice of Linacre and Wright (1992, pp. 58-59), individual responses were examined for patterns of misfit. Different patterns of misfitting responses were evident; examples are provided in Table 24 for illustrative purposes. The first row contains a misfit response for examinee 7. This examinee was given a rating of

2 (category 2) by rater 20 on the Thought and Detail (TD) scale for the Functional Writing (func) section. The expected rating (Expect) of 4.3 is based on the sum of the modeled probabilities of response to the various categories weighted by the category values (Wright & Masters, 1982, p. 97). The difference between the observed value and this expected value results in a residual of -2.3 (Residual), which has a studentized residual (t) of -3.

Examinee 128 illustrated the most common pattern of major misfitting responses for a misfitting individual. As shown, examinees like examinee 128 did not attempt a section, usually the third section, Response to Visual Communication. All three raters reported that the examinee had not completed this section.

Examinee 430 had an unusual response pattern: one rater awarded only 1s on the Personal Response to Literature scales, but, in contrast, only 5s on the Response to Visual Communication scales. In the case of examinee 671, one rater awarded 0s on the Personal Response to Literature scales while the second rater awarded a 5 on one Personal Response to Literature scale and the third rater awarded 5s on three of the Personal Response to Literature scales. Examinee 671's awarded marks can only be due to rater differences; in contrast, examinee 430's marks may be due to actual performance level differences on the two sections. However, the misfitting rater may have been more extreme in rating differences that were perceived by the other two raters.

Altogether, 40.0% of the total misfits were attributable to 0s being given to an examinee on one of the three sections with much higher marks being given on the scales of the other two sections. It would be reasonable to expect misfit in a model that has an assumption of a unidimensional trait (functional writing proficiency) when some scores resulted that are indicative of a second trait (lack of effort or perhaps poor time management). Since this second trait was observed in only 0.7% of the total number of

all score points given, it was judged to be no threat to the assumption of unidimensionality, nor was it judged to be a threat to the assumption of nonspeededness.

Table 24
Misfitting Responses

Examinee	Rater	Cat	Expect	Residual	t	Scale
7	20	2	4.3	-2.3	-3	funcTD
10	66	5	3.1	1.9	3	funcWRS
12	42	1	3.0	-2.0	-3	persORG
14	69	5	2.6	2.4	3	persMCO
25	66	5	2.9	2.1	3	persMCO
128	70	0	2.2	-2.2	-3	visTD
128	70	0	2.2	-2.2	-3	visORG
128	70	0	2.3	-2.3	-3	visWRS
128	69	5	2.4	2.6	3	persMCO
128	69	0	2.2	-2.2	-3	visTD
128	69	0	2.2	-2.2	-3	visORG
128	69	0	2.3	-2.3	-3	visWRS
128	26	0	2.4	-2.4	-3	visTD
128	26	0	2.3	-2.3	-3	visORG
128	26	0	2.4	-2.4	-3	visWRS
430	42	1	3.0	-2.0	-3	persTD
430	42	1	3.1	-2.1	-3	persORG
430	42	1	3.1	-2.1	-3	persMCH
430	42	1	3.1	-2.1	-3	persMCO
430	42	5	2.9	2.1	3	visTD
430	42	5	2.8	2.2	3	visORG
430	42	5	3.0	2.0	3	visWRS
671	43	0	2.5	-2.5	-3	persTD
671	43	0	2.6	-2.6	-4	persORG
671	43	0	2.7	-2.7	-4	persMCH
671	43	0	2.6	-2.6	-3	persMCO
671	60	5	2.9	2.1	3	persMCO
671	25	5	2.8	2.2	3	persORG
671	25	5	2.9	2.1	3	persMCH
671	25	5	2.9	2.1	3	persMCO
671	25	5	2.7	2.3	3	visTD
671	25	5	2.6	2.4	3	visORG
671	25	5	2.8	2.2	3	visWRS

Misfitting responses were also categorized by scale and by scale point awarded. Neither analysis revealed any indication of misfit pattern that would have suggested a misfitting scale or scale point.

Summary of Rasch Preliminary Analyses

The results of the four preliminary analyses indicated that the data were sufficiently unidimensional for a Rasch analysis, the scales were equally discriminating, there was no evidence of speededness, and the number of misfit responses was less than the number that would have indicated misfit of the data to the model. In summary, the data fit the Rasch model.

Multifaceted Rasch Analyses and Results

Rasch All Facet Summary

The Rasch analysis allowed the comparison of the three different facets on a common logit scale. The comparison of these facets is presented in Figure 4. The common logit scale (measr) is located on the extreme left hand side of the figure. The column label "+examinee" indicates that the more positive the value, the more proficient the examinee while the column labels "-rater" and "-scale" indicate that these scales have been reversed. Rater and scale (item) facets are reversed as severe raters and difficult items lower, rather than raise, examinee scores. For rater and scale facets, each asterisk represents one member; for the examinee facet, each asterisk represents 40 members. On the extreme right of the scale, in the "Point" column, the scale points are placed at the appropriate difficulty level on the logit scale. Vertical bars indicate increments of 0.25 logits.

Measr	+examinee	-rater	-scale	Point
7 +	+		+	+ (5) +
	.			
	.			
6 +	+		+	+ +
	.			
	.			
	.			
5 +	+		+	+ +
	.			
	.			---
	. more			
4 +	+ proficient	+	+	+ +
	.			
	*			
	*			4
3 +	+		+	+ +
	**			
	***	severe		
	****			---
2 +	+		+	+ +
	+			
	+			
	+			
1 +	+ *		+	+ 3 +
	+			
	+			
	+			
	+			
0 +	+		+	+ --- +
	+			
	+			
	+			
-1 +	+ *	lenient	+	+ +
	+			
	+			
	+			2
-2 +	+		+	+ +
	.			---
	.			
	. less			
-3 +	+ proficient	+	+	+ 1 +
	.			
	.			
-4 +	+		+	+ (0) +

Logit	* = 40 , < 40	* = 1 rater	* = 1	Point

Figure 4.
All facet summary.

As tables of more exact rater and scale characteristics are presented later in the chapter, rater number and scale names have been omitted for the sake of clarity and compactness of the figure. Statistics related to these facets are discussed in the appropriate sections following the interpretation of Figure 4.

The placement of all members of all facets on a common logit scale allows for the subjective comparison of facets on characteristics of central tendency, dispersion, and skewness. As seen in Figure 4, the examinees were relatively proficient as the center of their frequency distribution has a value greater than zero. The examinees ranged from approximately -3.50 to 6.75 logits. The large variation among examinees can be subjectively judged from the width of the histogram of examinee proficiencies found within Figure 4. The examinees' frequency distribution was seen to be slightly positively skewed. Kurtosis was not evident from the histogram.

The mean severity of the raters was set to zero. The raters were more uniform as a facet than examinees, varying from -1.00 to 1.00. This was desirable and expected as raters were trained to mark to a common standard whereas examinees were assessed as to their degrees of competency. Examinees were expected to vary in spite of teachers' attempts to educate them all, while raters were all trained to respond as uniformly as possible. The question remains as to whether the logit range of approximately -1 to 1 is so large that the raters cannot be considered a set of uniform members of the rater facet. Inspection of the histogram suggests that within this range, 9 raters were more severe than the main group of 53 raters while 8 raters were more lenient than the main group. Unlike the examinee histogram, the shape of the rater histogram suggests that the rater distribution was not skewed.

The mean difficulty of the scales was set to zero. The distribution of scale means, ranging from -0.25 to 0.50 logits, suggested the scales were of relatively uniform

difficulty. The scale histogram suggests 2, possibly 3, scales were slightly more difficult than the remaining 6 scales. When compared to the examinee frequency distribution, the spread of scale point values of 2 to 4 illustrated that essentially all examinees were assigned scale points by each rater in the range of 2 to 4 to match their proficiency level. The scale points of 0, 1, and 5 were used to describe relatively few examinees.

Comparisons with other studies The results presented to this point are similar to the results reported by Twing and Williams (1992) and more recently by Engelhard Jr. (1994) and Du (1995). Twing and Williams reported an examinee range of -9 to +7 logits, the examinees in Engelhard Jr.'s study ranged from -7.12 to +7.60, and the examinees in Du's study ranged from -4.5 to +10.0. Twing and Williams found a rater range of -0.64 to +0.64 logits while the raters in Engelhard Jr.'s study ranged from -1.22 to +1.12, and in Du's study from -2 to +2. The two marking scales used in the Twing and Williams' study ranged in difficulty from -0.95 to +0.95 while the scales in Engelhard Jr.'s study ranged to -0.51 to 0.48, and the scales in Du's ranged from -1.0 to +0.5. Taken together, the results of the three studies and the present study are quite similar and would lead to the same interpretation.

Rater Characteristics

Rater Rasch Severity

The rater severity characteristics are presented in Table 25. The results are arranged in order of logit severity, beginning with the most severe rater. Rater 58 had an observed average of 2.5; this is the rater mean of the observed total scores awarded to the examinees marked by that rater divided by nine, the number of scales. Rater 58's fair average was 2.2. As described in Chapter III, this statistic is a linear transformation of the logit severity, 1.03. The model error for rater 58 was 0.05.

Table 25
Rater Rasch Severity

Rater	Observed Average	Fair Average	Severity logit	Model Error
58	2.5	2.2	1.03	0.05
5	2.7	2.3	0.82	0.04
57	2.7	2.4	0.78	0.04
36	2.6	2.4	0.75	0.04
59	2.7	2.4	0.59	0.03
68	2.7	2.4	0.58	0.03
1	2.8	2.4	0.56	0.04
52	2.7	2.5	0.52	0.04
50	2.8	2.5	0.48	0.04
65	2.8	2.5	0.37	0.04
21	2.8	2.5	0.36	0.04
31	2.8	2.5	0.35	0.03
10	2.8	2.5	0.33	0.03
70	2.8	2.6	0.31	0.04
56	2.9	2.6	0.22	0.03
54	2.9	2.6	0.22	0.03
30	2.8	2.6	0.21	0.04
23	2.9	2.6	0.20	0.03
69	2.9	2.6	0.19	0.04
67	2.9	2.6	0.18	0.03
37	2.9	2.6	0.15	0.05
32	2.8	2.6	0.14	0.04
35	2.9	2.6	0.14	0.03
49	2.9	2.6	0.14	0.03
38	2.9	2.6	0.13	0.04
43	2.9	2.6	0.11	0.03
19	2.9	2.6	0.11	0.04
16	2.9	2.6	0.07	0.04
9	2.9	2.6	0.06	0.03
17	3.0	2.6	0.06	0.04
20	2.9	2.6	0.06	0.03
48	2.9	2.7	0.05	0.02
3	2.9	2.7	0.02	0.04
2	3.0	2.7	-0.03	0.04
41	2.9	2.7	-0.03	0.04
51	3.0	2.7	-0.05	0.04
33	2.9	2.7	-0.05	0.04
4	3.0	2.7	-0.05	0.03
34	3.1	2.7	-0.07	0.04
11	3.0	2.7	-0.09	0.03
39	3.0	2.7	-0.09	0.03
28	3.1	2.7	-0.09	0.04
47	2.9	2.7	-0.11	0.06
26	3.0	2.7	-0.12	0.03

Rater	Observed Average	Fair Average	Severity logit	Model Error
6	3.0	2.7	-0.13	0.04
15	3.0	2.7	-0.15	0.03
14	2.9	2.7	-0.15	0.05
18	2.9	2.7	-0.17	0.04
55	3.0	2.7	-0.18	0.03
63	3.0	2.7	-0.19	0.03
8	3.0	2.7	-0.20	0.03
45	3.0	2.7	-0.20	0.03
29	3.0	2.7	-0.20	0.04
44	3.0	2.8	-0.21	0.03
62	3.1	2.8	-0.22	0.04
7	3.0	2.7	-0.22	0.03
13	3.1	2.8	-0.27	0.04
24	3.1	2.8	-0.29	0.04
60	3.1	2.8	-0.29	0.05
66	3.0	2.8	-0.32	0.04
25	3.1	2.8	-0.33	0.04
61	3.1	2.8	-0.36	0.03
12	3.2	2.8	-0.44	0.04
64	3.2	2.9	-0.55	0.04
40	3.2	2.8	-0.56	0.03
53	3.2	2.9	-0.67	0.04
27	3.3	2.9	-0.69	0.05
46	3.3	3.0	-0.80	0.04
42	3.3	3.0	-0.87	0.03
22	3.3	3.1	-1.04	0.03
Rater	Observed Average	Fair Average	Severity logit	Model Error
Mean	2.9	2.7	0.00	0.04
SD	0.2	0.2	0.39	0.01

As shown in Table 25, the observed average severity for the raters ranged from 2.5 to 3.3 with a mean of 2.9 and a standard deviation of 0.2. The fair average severity for the raters ranged from 2.2 to 3.1 with a mean of 2.7 and a standard deviation of 0.2. The logit severity for raters ranged from 1.03, the most severe, to -1.04, the most lenient. The mean logit severity was fixed at 0; the standard deviation of logit severity was 0.39. The mean model error, RMSE, was 0.04.

Rater facet severity statistics. The rater severities were distributed normally, (normal chi-square: $\chi^2_{67} = 68.96$, $p < .41$), but displayed significant differences, (fixed

chi-square: $\chi^2_{69} = 7696.81$, $p < .001$). Other Rasch group indices concur with the results of the fixed chi square test. The reliability of separation of raters was 0.99 and the interrater reliability, IRR, was .01. The separation index was 10.30 and the strata index was 14.07. The values of the reliability of separation index and the IRR indicate that the raters were reliably separated and ordered based on severity; the separation index of 10.30 means the adjusted standard deviation was more than 10 times the RMSE; the strata index states there are 14 distinct levels of raters based on severity. These results are similar to those reported by Engelhard Jr. (1994). He found a significant fixed chi square ($\chi^2_{15} = 170.7$, $p < .01$) and a reliability of separation index of .87.

When there is evidence of rater severity within the group of raters, it is appropriate to identify the individual raters who display severe and lenient scoring frequencies (see Chapter IV, pp. 79-81). However, there is a lack of guidelines for distinguishing a severe or lenient rater from a rater that displays little difference from his or her peers. Within a Rasch analysis there would be little emphasis placed on identification of raters who differ in severity as the logit score for an examinee is calculated taking into account the severity of the rater. However, as Zegers (1991) has previously pointed out, doubt is cast upon a judgment procedure in which there is a large amount of discrepancy (p. 321). For this reason, procedures for distinguishing between rater severities are proposed.

Three methods for distinguishing severe and lenient raters were considered in the present study. First, as discussed during the description of the content of Figure 4, the histogram of logit severities was examined to identify sub groups of raters that were seen to be more lenient or more severe than the main sub group of raters. This subjective approach suggested subgroups of 9 severe and 8 lenient raters.

To help make this approach more "objective", the suggestion of Sax (1984) that natural breaks in the distribution of the scores be used to establish cut off scores was applied to the logit severities. Cut off scores were established as follows: the first cut off score was placed above the mean severity between the first pair of adjacent rater severities which differed by at least 2.5 times the model error; the second cut off score was placed below the mean between the corresponding pair of severities. In the present case, this minimum value of the natural break was .10. Employing those cut offs, 9 raters were identified as severe and 7 raters were identified as lenient.

As was the case for classical test score theory, an analytical approach was also considered based upon the multiple t approach: confidence intervals were constructed around each rater's logit value. The standard error used was the model fit for the rater ; this value was multiplied by 2.5 to produce the upper and lower limits of each rater's confidence interval. Using this approach, 27 raters were identified as severe and 31 raters were identified as lenient.

Rater Consistency

Rater consistency refers to the extent which the observed scores awarded by a rater agree with the expected scores to be awarded by that rater according to the prediction of the Rasch model. In Table 26 the raters are sorted by their infit mean square statistic values and then by their respective t statistic values within bands of equal infit values. The first thing to note is the similarity of the corresponding infit and outfit statistics. Both fit statistics ranged from 0.7 to 1.5, with the mean values of 1.0 and standard deviation values of 0.2. The two t statistic values ranged from -9 to 9 with means of approximately 0 and standard deviations of 5.2.

Table 26
Rater Mean Fit Statistics

Rater	Infit Mean Square	t	Outfit Mean Square	t
5	0.7	-9	0.7	-9
54	0.7	-9	0.7	-9
32	0.7	-9	0.7	-9
57	0.7	-7	0.7	-8
4	0.8	-8	0.8	-8
35	0.8	-7	0.8	-7
9	0.8	-7	0.8	-8
30	0.8	-6	0.7	-6
58	0.8	-5	0.8	-5
59	0.8	-5	0.8	-5
1	0.8	-4	0.8	-4
52	0.8	-4	0.8	-4
29	0.8	-4	0.8	-4
62	0.8	-4	0.8	-5
60	0.8	-4	0.8	-4
15	0.9	-4	0.9	-4
25	0.9	-4	0.9	-4
40	0.9	-4	0.9	-4
31	0.9	-3	0.9	-3
43	0.9	-3	0.9	-4
19	0.9	-3	0.9	-3
63	0.9	-3	0.9	-3
12	0.9	-3	0.9	-3
46	0.9	-3	0.9	-3
70	0.9	-2	0.9	-2
67	0.9	-2	0.9	-1
49	0.9	-2	0.9	-2
41	0.9	-2	0.9	-2
39	0.9	-2	0.9	-2
6	0.9	-2	0.9	-2
8	0.9	-2	0.9	-2
36	0.9	-1	1.0	0
47	0.9	-1	0.9	-1
37	1.0	-1	0.9	-1
20	1.0	-1	1.0	-1
48	1.0	-1	1.0	-1
26	1.0	-1	0.9	-1
7	1.0	-1	1.0	-1
50	1.0	0	1.0	0
23	1.0	0	1.0	0
17	1.0	0	1.0	0
2	1.0	0	1.0	0
11	1.0	0	1.0	0
61	1.0	0	1.0	-1

Rater	Infit Mean Square	t	Outfit Mean Square	t
22	1.0	0	1.0	-1
68	1.0	1	1.0	1
56	1.0	1	1.0	1
13	1.1	1	1.1	1
14	1.1	2	1.1	2
21	1.1	3	1.1	3
16	1.1	3	1.1	4
45	1.1	3	1.1	3
44	1.1	4	1.1	3
51	1.2	5	1.2	5
64	1.2	5	1.2	5
10	1.2	6	1.2	5
55	1.2	7	1.2	7
42	1.2	7	1.2	7
65	1.3	6	1.3	6
69	1.3	6	1.2	5
3	1.3	6	1.3	6
33	1.3	6	1.3	6
24	1.3	6	1.3	6
34	1.3	7	1.3	7
53	1.3	7	1.3	7
28	1.3	9	1.3	8
66	1.4	8	1.4	8
38	1.4	9	1.4	9
18	1.4	9	1.4	9
27	1.5	8	1.5	8
Rater	Infit Mean Square	t	Outfit Mean Square	t
Mean	1.0	-0.2	1.0	-0.3
SD	0.2	5.2	0.2	5.2

In a MFRM analysis, an omnibus test of differences of fit mean square values among raters is not performed. Judgment about the raters as group is made on the basis of the proportion of raters who fall outside of accepted guidelines. One guideline, the standardized t-statistic, was intended to provide an indication of statistical significance of the fit statistic and, as its name suggests, it was designed to be interpreted as a Student t value (Smith, 1995). As pointed out in Chapter III (see p. 58), this t-statistic is known to give highly inflated values for large sample sizes (more correctly, high counts per facet

member). Therefore agreed upon acceptable limits for fit values are also, or often alternately, employed.

When the test of the Rasch $\text{inf}|\text{t}| > 2$ was employed, 24 raters were found to be too constrained and 21 raters found to be too erratic. An erratic rater has a tendency to commonly award scores that were different than what was expected for that rater (Wright & Linacre, 1994, p. 370). These results have been shaded in Table 26. Using the guidelines for raters of $0.4 < \text{FitMS} < 1.2$, indicated with a double line in Table 26, no rater was found to be too constrained, that is, no rater had a tendency to continually award the same point across examinees. However, 17 raters were found to be too erratic. The two methods for judging the raters differed in both sensitivity and skewness. Given that 24.3% of the raters were considered too erratic based on the FitMS guidelines and 64.3% of the raters were different based on the t statistic, it appears that raters were not homogeneous in terms of rater consistency.

Rater Agreement

Rater agreement was assessed using the rater Rasch point biserial. Table 27 presents the rater Rasch point biserials. The raters are ordered from highest to lowest. The Rasch point biserial values ranged from .36 to .50, with a mean of .43 and a standard deviation of .03.

Given the relative recency of application of Rasch agreement statistics to raters, there appear to be no guidelines in the literature for interpretation of rater Rasch point biserials. If rater point biserials are treated like rater intercorrelations, no rater appeared to be in consistent agreement with his or her group of fellow raters as none fell within 2 standard deviations of unity. But if the Rasch point biserial is treated like a classical item point biserials then rater point biserials greater than the limit $.00 + 2\sigma_r$ (Crocker & Algina, p. 386) would be identified as agreeing. Inspection of the results in Table 27 reveals all

raters are acceptable. In order to better interpret this statistic, the rater point biserials are discussed further after the discussion of scale and examinee point biserials.

Table 27
Rasch Rater Agreement

Rater	Rasch Point Biserial	Rater	Rasch Point Biserial	Rater	Rasch Point Biserial
47	0.50	15	0.45	6	0.41
31	0.49	55	0.45	61	0.41
56	0.49	62	0.45	53	0.41
29	0.49	7	0.45	42	0.41
52	0.47	27	0.45	10	0.40
50	0.47	68	0.44	39	0.40
19	0.47	20	0.44	26	0.40
16	0.47	36	0.43	66	0.40
4	0.47	32	0.43	58	0.39
24	0.47	43	0.43	1	0.39
25	0.47	8	0.43	3	0.39
64	0.47	44	0.43	2	0.39
59	0.46	60	0.43	28	0.39
41	0.46	30	0.42	45	0.39
14	0.46	23	0.42	46	0.39
63	0.46	35	0.42	21	0.38
13	0.46	33	0.42	69	0.38
70	0.45	18	0.42	9	0.38
54	0.45	22	0.42	12	0.38
37	0.45	5	0.41	40	0.38
49	0.45	57	0.41	34	0.37
17	0.45	67	0.41	65	0.36
48	0.45	51	0.41	38	0.36
11	0.45				
Rater	Rasch Point Biserial	Rater	Rasch Point Biserial	Rater	Rasch Point Biserial
Mean					0.44
SD					0.03

Rasch Measurement of Scales

Rasch Scale Difficulties

Table 28 contains scale difficulties, presented in order of decreasing difficulty. The means of the scales were more similar in difficulty than the raters were similar in severity (see Figure 4). Three scales, measuring technical aspects of writing --

persMCO, persMCH, and funcWRS-- were generally the easiest. The Thought and Detail scale for the Functional Writing, funcTD, was the exception, being the easiest scale with a difficulty of -0.30 logits.

Table 28
Scale Difficulties

Scale Name	Observed Average	Fair Average	Difficulty Logit	Model Error
visORG	2.8	2.5	0.43	0.01
visTD	2.8	2.5	0.39	0.01
visWRS	2.9	2.6	0.20	0.01
persTD	2.9	2.7	0.02	0.01
persORG	3.0	2.7	-0.12	0.01
funcWRS	3.0	2.7	-0.14	0.01
persMCO	3.0	2.8	-0.23	0.01
persMCH	3.1	2.8	-0.25	0.01
funcTD	3.1	2.8	-0.30	0.01
Mean	3.0	2.7	0.00	0.01
SD	0.1	0.1	0.26	0.00

Scale difficulty statistics. The scale difficulties were distributed normally, (normal: $\chi^2_6 = 8.00$, $p < .24$) but displayed significant differences (fixed: $\chi^2_8 = 3792.57$, $p < .001$). Other Rasch group indices concur with the results of the fixed chi square test. The reliability of separation of scales was 1.00, the separation index was 20.43 and the strata index was 27.57. The reliability of separation index of 1.00 indicates that the scales were perfectly separated and ordered based on difficulty; the value of the separation index, 10.30 meant the adjusted standard deviation was more than 10 times the RMSE, while the strata index indicated there were 27 distinct levels of difficulty within the range of scale difficulties. All indices that were part of a Rasch analysis all gave the same result: scales differed as elements and as a group.

The statistics just presented appear to be a contradiction to the visual image presented in the all facet summary, Figure 4. There, the scale was the most tightly grouped facet while the members of the examinees facet were spread far apart. When the fair averages are converted to a ratio of observed average to maximum scale value,

analogous to classical item p values, the values ranged from $p = .50$ to $p = .56$ on the visORG to funcTD respectively, suggesting the scales were quite similar in difficulty. In contrast, the separation index indicated that the scales were much more different from each other than the raters were from other raters, or the examinees were from other examinees (see the examinee discussion which follows the scale discussion). The cause of this seeming contradiction is the very small value of the RMSE, 0.01. The reliability of separation, the separation index, and the strata index will have large values if the RMSE is small. Given, the very small RMSE, therefore, scales will be found to be different, in spite of little or no differences among scales in a practical sense.

Scale Consistency

The second characteristic on which members of a scale facet may differ is the way in which scale points are used in comparison with what is expected as measured by the fit mean square statistic. The results for scale are presented in Table 29 in order of increasing infit values. Again as with raters, the corresponding infit and outfit measures are very similar. As shown by these values, the technical writing scales were slightly more constrained than the others; the persMCH was an exception. The TD scales were the most erratic. However, the infit values only ranged from 0.8 to 1.1. Guidelines of $0.6 < \text{FitMS} < 1.5$ suggested no scales were marked too erratically nor was any scale marked too uniformly.

As with raters, the large numbers of observations for each scale produced a t statistic that resulted in values greater than the absolute value of 2; visORG was the exception here. Again, conflicting interpretations exist. The guidelines approach claimed no scales were too constrained or too erratic while the t statistic approach claimed all scales but one misfit in one manner or the other.

Table 29
Scale Mean Square Fit Statistics

Scale Name	Infit		Outfit	
	Mean Square	t	Mean Square	t
persMCH	0.8	-9	0.8	-9
visWRS	0.9	-9	0.9	-9
funcWRS	0.9	-7	0.9	-8
persORG	0.9	-5	0.9	-5
visORG	1.0	2	1.0	1
persMCO	1.1	4	1.1	4
visTD	1.1	7	1.1	6
persTD	1.1	9	1.1	9
funcTD	1.1	9	1.1	9
Mean	1.0	0.2	1.0	-0.1
SD	0.1	7.4	0.1	7.4

Scale Agreement

Table 30 is ordered by decreasing Rasch point biserial. When ordered by the Rasch point biserials, the scales perfectly followed the order: technical scales first, followed by the ORG scales, and finally the TD scales. These Rasch point biserial values ranged from .38 to .50, with a mean of .44 and a standard deviation of .03.

Table 30
Scale Rasch Point Biserial

Scale Name	Rasch Point Biserial
visWRS	0.50
persMCH	0.47
funcWRS	0.46
persMCO	0.46
persORG	0.45
visORG	0.44
persTD	0.43
visTD	0.41
funcTD	0.38
Mean	0.44
SD	0.03

As with classical item intercorrelation, high values are more desirable. If a confidence band of 2 SD less than unity is produced, no scale was found within that interval. If the minimum acceptable classical point biserial of $.00 + 2\sigma_p$ (Crocker &

Algina, 1986, p. 326) were applied, even with σ_p taken as SD, all scales meet this minimum criterion.

Rasch Measurement of Examinees

Examinee Proficiency (Level of Achievement)

Just as classical analyses resulted in information about the examinees used in this study, the description of the Rasch analyses included an examination of the examinees as well. Given the large number of examinees, 4927, retained in this analysis, only a summary of examinee characteristics is presented here in Table 31. Unlike rater severity and scale difficulty, examinee mean proficiency was not centered at 0 logits. As both rater mean severity and scale mean difficulty were centered at 0 logits the mean fair average should be equal to the mean observed average. As the observed average was more accurately known to be 2.95, and the fair average known to be 2.94, the different reported values of 3.0 and 2.9 were due to rounding error.

Table 31
Examinee Characteristics Summary

Examinee	Obs Ave	Fair Ave	Logit	Mod Err	Infit		Outfit		RPt Bis
					MSq	t	MSq	t	
Mean	3.0	2.9	0.76	0.30	1.0	-0.3	1.0	-0.3	0.15
SD	0.5	0.5	1.31	0.02	0.7	1.7	0.7	1.7	0.20

Examinee facet proficiency. The examinee proficiencies were distributed normally (normal $\chi^2_{4924} = 4,943.92$, $p < .42$) but displayed significant differences, (fixed $\chi^2_{4928} = 105,839.50$, $p < .005$). Other Rasch group indices concur with the results of the fixed chi square test. The reliability of separation of examinees was 0.95; the separation index was 4.22 and the strata index was 5.96. The reliability of separation index of .95 indicated that the examinees were reliably separated and ordered based on severity, the separation index of 4.22 meant the adjusted standard deviation was more

than 4 times the RMSE, while the strata index indicated there were 6 distinct levels of examinees based on proficiency.

Examinee Consistency

Again the infit and outfit values were very similar; both had a mean of 1.0 and a standard deviation of 0.7. When compared to raters and scales, the examinees were more variable with infit standard deviations of 0.7, in contrast to raters, 0.2, and scales, 0.1 (see Table 26 and Table 29 respectively). If the distribution of examinee infit values is not badly skewed, then the relatively large standard deviation for the infit statistic suggests the examinees displayed relatively more instances of low infit values (overly-consistent) and more instances of high infit values (erratic) than did the raters or the scales. This is not unexpected as some examinees would do equally well on all parts of the examination, others would be more variable (see discussion of examinee 128, p. 121). Such variation reflects the individual differences that are usually found among students. The raters are seen to be behaving in a much more consistent fashion, and in a fashion similar to the scales.

Interpretation of Rasch Point Biserials

The most striking difference between the examinee facet and the other two facets was the low Rasch point biserial. The very low mean value, .15, coupled with a relatively large standard deviation, .20, was considered reasonable for examinees, as while examinees as a group were educated to perform well as mentioned above, it was not expected that they would all perform in the same fashion.

The higher mean and lower standard deviation of the rater Rasch point biserial, .43 and .03 respectively, were essentially identical to the mean and standard deviation for the scale Rasch point biserial, .44 and .03 respectively. According to this statistic the raters behaved essentially like the scales and not like the examinees, a desirable quality

for raters as the raters are part of the measurement process, not the object of measurement.

Correction of Examinee Scores for Rater Effects

The Rasch analysis produced a fair average which is a linear transformation from the Rasch logit score to an observed score metric (see p. 65). The distribution of point difference frequencies is presented in Table 32. As shown, the differences between the observed score and the corresponding fair average score ranged from -3 to 3. Since both the rater facet and scale facet were centered on zero, the mean examinee fair average and the examinee observed average were expected to be the same. The difference between the mean observed score and mean fair average score, 0.03 points, is attributable to rounding error.

Table 32
Differences Between Rasch Fair Averages and Observed Scores

Point Difference	Percent in Interval
-3	0.1%
-2	3.0%
-1	21.9%
0	48.9%
1	21.4%
2	4.3%
3	0.3%

Just over 51% of the examinees received corrected scores that differed by one or more points, while approximately 8% received marks that differed by two or more points. The standard deviation of the difference between the corresponding fair average and observed average scores, the CRMS, was 0.84 while the average absolute difference

(AAD) was 0.66. Both the CRMS and the AAD corresponded with what is shown in Table 32; most of the corrected scores did not vary by more than one point from the uncorrected score and virtually all the corrected scores were within two points of the uncorrected score.

The correlation between the logit score, and the observed score was .99. The correlation between the logit score, and the multiple choice score was only slightly higher at .56 than the correlation between observed score and the multiple choice, .55. This minimal increase in correlation between a logit measure and observed score measure was also noted by Twing and Williams (1992, p. 8). The increase in correlation was expected to be of this magnitude given the increases in correlation reported by Braun (1988, p. 93).

The range of differences between the observed score and the fair average score is of some concern. As indicated previously, differences of one point or greater were considered to be important (see p. 15). The one point difference for over half the examinee population should not be overlooked. A score difference that is unacceptable for one portion of the examination cannot be acceptable for another portion of the examination. As was the case for the linear correction, in light of the *Principles for Fair Student Assessment Practices for Education in Canada* , the use of a correction formula is justified so that all examinees receive fair and equitable treatment (p. 3).

CHAPTER VII COMPARISONS AMONG APPROACHES

This chapter consists of a series of comparisons among the three approaches that were the focus of this study. The first research question of this study was concerned with the detection of rater variability that is known to exist among the raters. The second concerned corrections that can be applied to correct for the influence of this variation on examinees scores. The comparisons are organized in terms of these two questions.

In the first section, comparisons are ordered in terms of the characteristics of rater variation considered. The order is rater severity, followed by rater consistency and then rater agreement. For each characteristic, a discussion of the characteristics applied to scales is presented first to assist with the interpretation of the corresponding comparisons for rater. In the second section, comparisons were made between the two corrections considered in this study: a classical test theory linear scaling and the multifaceted Rasch fair average. A linear regression approach was also initially considered but was rejected as unfeasible given the relatively empty data matrix.

Comparison of Detection Approaches

Rater Severity

As pointed out at various points in Chapter VI, the MFRM is applied to all facets considered in the analysis. In the case of severity, scale severity is better labeled difficulty in that it corresponds with what is found in classical test score theory analysis. Indeed, the correlation between the observed scale difficulties and the logit difficulties was 1.00, a correlation value that is very similar to the Twing and Williams (1992) result of .98. This correlation suggested that the classical scale difficulty and the Rasch logit difficulty for the scales ranked the scales identically. However, the variables may correlate highly, yet be systematically different. To address the issue, the logit severity was rescaled for each

scale. This fair average is in the metric of difficulty in the classical score framework. Taking account of the difference of the mean observed average score and the mean fair average scale score of 0.274 (see p. 125), the residuals between the observed average score and the corresponding fair average score were all 0.04 or less and likely explainable as rounding error in the fair average scores. Thus in terms of the correlation and the absolute deviation the scale difficulties yielded by the models were essentially the same.

Turning now to raters, the results for each of the three approaches are presented separately in Table 33. Some values are given for purposes of completeness but are not discussed in the comparisons, for example, the range of logit values. Other results related to comparisons are not placed in the table as they are the result of additional analysis in which the approaches are compared. For example, as illustrated in the previous paragraph, correlations and absolute difference between "scores" yielded by two approaches are considered.

Omnibus Results

Inspection of the omnibus results reveals that the Alexander-Govern A statistic and the MFRM fixed chi square and related Rasch statistics revealed variation in rater severity. The variance component for raters yielded in the generalizability analysis suggested that variation in raters was low. Thus it would appear that the three approaches are differentially sensitive to rater variation, with CTT and MFRM agreeing and G theory not.

Table 33
Comparison of Severity Differences

Classical Test Theory	Generalizability Theory	Multifaceted Rasch Model
Scale		
observed scale difficulties p values = .56 – .62, no omnibus test of differences performed	$\sigma_i^2 = 0.82\%$ $\sigma_i = 0.077$	scale logit difficulties range -0.30 – 0.43, scale fair averages range of p values = .50 – .56, $\chi_i^2 = 3792.57$, $p < .001$
Rater		
AG A statistic = 928.51, p < .001 rater severity scores range 22.88 – 30.02 AG z _j , range -9.634 – 7.534	$\sigma_i^2 = 4.76\%$ $\sigma_i = 0.187$ individual severities would be the classical severities	fixed $\chi_i^2 = 7696.81$, p < .001 reliability of separation of raters = .99 separation index = 10.30 strata = 14.07 IRR = .01 rater logit severities range -1.04 – 1.03 rater fair averages range 2.2 – 3.1
AG z _j , .05 level 18 severe raters, 19 lenient AG z _j , .01 level 14 severe raters, 15 lenient		model error, .01 level 27 severe raters, 31 lenient histogram 9 severe raters, 8 lenient natural breaks 9 severe raters, 7 lenient
of 10 most severe, 9 identified by Rasch severity of 10 most lenient, 9 identified by Rasch severity		of 10 most severe, 9 identified by AG z _j of 10 most lenient, 9 identified by AG z _j

Identification of Raters Displaying Severity

Given the omnibus results for the classical and Rasch approach, further analyses were conducted within each framework to identify raters that were severe or lenient. Comparisons of the two sets of results revealed a mixed result. The classical AG z score identified 37 raters as significantly different (18 severe and 19 lenient) at the .05 level of significance, and 29 raters as significantly different (14 severe and 15 lenient) at the .01 level of significance. The analogous Rasch model error procedure identified 58 raters as significantly different (27 severe and 31 lenient) at the .01 level of significance. Within the Rasch approach, two other procedures were employed in this study for identifying raters from the logit severity distribution: judgment analysis of the histogram and the natural breaks approach. These results were comparable. The histogram approach identified 9 severe and 8 lenient while the natural breaks approach identified 9 severe and 7 lenient. Application of the natural breaks procedure to the classical procedure resulted in at best one rater being identified as severe. Clearly these results are very different from the analytic AG z and model error results.

To examine further the discrepancies noted between the two approaches, the correlation between the rater classical severities and the Rasch logit severities was examined. Its value, $-.96$, suggested a strong agreement in the ranking of the raters. Further, when the ten most severe raters were identified using the classical AG z score, nine of the ten were also found within the ten most severe by Rasch standards. And nine of the ten most severe raters identified using Rasch logit severity were among the ten most severe by classical measures. This procedure was repeated for the ten most lenient raters. Again, nine out of the ten identified by one procedure were identified using the other. The descriptive statistics for the four raters that were not identified by both procedures were examined. A combination of relatively high standard deviation and

lower than average numbers of papers marked caused raters to not be identified by the AG z score. The opposite combination of relatively low standard deviation and higher than average numbers of papers favoured identification by AG z criteria.

As noted earlier, the percent variation due to raters as identified by G analysis was 4.76%. This indicates that while rater variation as a source of variation did not contribute greatly to variation among scores, it does not necessarily mean that there will not be some raters with significant deviations in severity. This interpretation is consistent with the histogram procedure and natural breaks procedure results from Rasch but inconsistent with the statistical conclusions of both the omnibus A test (Alexander and Govern) of the classical and the omnibus fixed chi-square employed in the Rasch. Thus, it would appear that the G theory results are consistent with the Rasch histogram and natural breaks results.

Rater Consistency

To add to the discussion of the comparison of CTD consistency measures for raters with the infit and outfit consistency index for raters, a brief comparison of scale measures is first presented. Unlike classical scale difficulty and scale logit difficulty which correlated almost perfectly, the correlation between the standard deviations of the scales and their Rasch infit statistics was only .76; the correlation between the standard deviation and the infit t statistic was .67. With 59% shared variance, it is apparent that the standard deviation was not measuring exactly the same characteristics as the fit statistics, yet there was a considerable overlap in the interpretation of the two variables.

Turning now to raters, the results related to rater consistency are presented in Table 34. Again some results related to comparisons are not placed in the table as they are the result of a comparison between approaches and so do not belong in any one column, for example, the correlation between CTD and infit.

Table 34
Comparison of Rater Consistency

Classical Test Theory	Generalizability Theory	Multifaceted Rasch Model
Rater		
BF test $F = 5.85$ $p < .0005$		
χ^2 test .05 level 16 central tendency bias, 13 not consistent raters χ^2 test .01 level 10 central tendency bias, 6 not consistent raters		Rasch t , .05 level 24 too consistent, 21 erratic FitMS rule 0 too consistent, 17 erratic
of 10 largest central tendency error, 7 identified by Rasch measure of 10 least consistent, 5 identified by Rasch measure		of 10 most consistent 7 identified by CTD of 10 most erratic, 5 identified by CTD

Classical Rater Consistency and Rasch Rater Consistency

The similarity in interpretations of rater CTD and rater infit and outfit statistics lead to their comparison. Popham's (1990) definition of central tendency and the modified definition used in the present study appear to be a subset of a response set that Engelhard Jr. (1994) refers to as "halo". The presence of a low CTD_j for a rater is indicative of central tendency. Engelhard Jr.'s halo is detected by low infit and outfit values.

As shown in Table 34, and as noted in Chapter V, central tendency cannot be detected in a G analysis. And, as shown in Table 34, and noted in Chapter VI, while consistency is addressed in the Rasch approach, there is no overall omnibus test. Instead, guidelines are used to identify raters with Fit mean squares or corresponding t values outside suggested limits. In the case of the classical approach, a test of homogeneity of rater CTD² can be conducted with the choice of test dependent on the

shape of the distribution of CTD_j^2 . In the case at hand, the Brown-Forsythe test was used; the results indicated that there were significant differences among the raters' CTDs.

In the classical approach, the chi square test for single variances was used to identify discrepant raters. Working at the .05 level of significance, 16 raters were identified as possessing central tendency error or using Rasch terminology, as being more constrained than the other raters. Thirteen were identified as less constrained. At the .01 level of significance, 10 raters were identified as more constrained and 6 as less constrained. When the Rasch t statistic test, $|t| > 2$, was used, 24 raters were identified as being more constrained and 21 as less constrained (erratic) raters. Using the Rasch rule of thumb for judged ratings, $0.4 \leq \text{FitMS} \leq 1.2$, no raters were considered too constrained and 17 raters were considered too erratic in ratings.

Unlike classical rater severity and rater logit severity which correlated almost perfectly, the correlation between the rater central tendency deviation and the Rasch infit statistic was only .75. Thus some differences between the numbers identified by the two procedures was expected. To assess this, the ten most extreme raters at each end of the distribution of raters using the CTD_j measure and the Rasch infit mean square measure were compared. In the case of over consistency, seven of the ten most consistent by CTD_j measures were also found within the ten most consistent by Rasch infit. This procedure was repeated for the ten most lenient raters. Here, five out of the ten identified by one procedure were identified using the other.

Rater Agreement

Scale test-score correlations and the Rasch scale biserials correlated .96 with each other. The meaning of the high correlation between these scale statistics was easily seen when the rank order of the scales was examined. The scales were in virtually identical rank order; the discrepancy from unity is attributable to the use of two decimal places in

the reporting of the Rasch point biserials. The mean of the scale test-score correlations adjusted for common scale was .66 with a standard deviation of .06. The mean of the Rasch point biserials was .44 with a standard deviation of .03. The two indices functioned similarly although the Rasch point biserial values were systematically lower and with a smaller standard deviation. Either index would yield the same interpretation for a relative decision but the two statistics cannot be compared according to their magnitudes.

Results related to rater agreement are presented in Table 35. Again some results related to comparisons are not placed in the table as they are the result of a comparison between approaches.

Table 35
Comparison of Rater Agreement

Classical Test Theory	Generalizability Theory	Multifaceted Rasch Model
Scale		
scale intercorrelations mean .66, SD .06		scale Rasch point biserials mean .44, SD .03
Rater		
rater intercorrelations mean .63, range .42 - .77		rater Rasch point biserials mean .44, range .36 - .50
CI spans .8 rule 6 satisfactory 2SD from 0 rule all satisfactory		CI spans .8 rule 0 satisfactory 2SD from 0 rule all satisfactory
mean interrater $r = .63$ $\alpha = .89$ Spearman-Brown = .84	Hoyt ANOVA = .90 Full analysis = .79	reliability of separation of examinees = .95

In direct contrast to the scale results, the correlation between the rater intercorrelations and the Rasch rater point biserials was .31. These two statistics are seen

to be measuring very little in common with only 9.6% shared variance. The interrater correlations ranged from .42 to .77, with a mean of .63. The mean of the Rasch rater point biserials ranged from .36 to .50, with a mean of .44.

Individual rater agreement indices were judged using several guidelines. The first guideline, an "agreement" index above .8, resulted in six raters being identified as satisfactory when classical test theory intercorrelations were used. When applied to the Rasch point biserials, this guideline resulted in no raters being seen as satisfactory. A second rule, an "agreement" index greater than 2 standard deviations above zero in the distribution of the index across raters resulted in all raters being judged as satisfactory using both of intercorrelations and the Rasch point biserials.

The Rasch point biserial appears to have the same purpose, namely to look at the "agreement" between an element of a facet and the remaining elements of that facet. Hence, it was expected the results from the application of both would be similar. This as noted above, was the case for scales but not for raters. The difference in findings may be an artifact of the design. More specifically all examinees are marked on all scales, all raters marked all scales but all raters do not encounter all examinees. This result was not investigated further.

Another approach to assessing the reliability of the scores produced is the calculation of test reliability statistics. As shown in Table 35, the classical coefficient α , computed taking into account scales, was .89. When the Spearman-Brown prophecy formula was applied to the mean interrater correlation, the result was .84. The generalizability coefficient taking into account raters and scales was .82. The classical coefficient α is greater than the generalizability coefficient because of the lack of recognition of raters as a facet in the design. The Rasch reliability of separation for examinees was .95.

Comparison of Corrections

The results of linear scaling, the multifaceted Rasch model correction, and the comparison between the two procedures are presented in Table 36. All correction procedures were applied to examinee scores, hence discussion of corrections is a discussion of examinee scores.

Table 36
Comparison of Correction Procedures

CTT Linear Scaling		Multifaceted Rasch Model		Comparison	
-3	0.2%	-3	0.1%	-3	0.0%
-2	3.4%	-2	3.0%	-2	0.3%
-1	22.4%	-1	21.9%	-1	13.7%
0	45.8%	0	48.9%	0	74.6%
1	24.7%	1	21.4%	1	10.9%
2	3.3%	2	4.3%	2	0.4%
3	0.0%	3	0.3%	3	0.0%
CRMS = 0.82 AAD = 0.65 correlation with uncorrected = .99 correlation with multiple choice = .56		CRMS = 0.84 AAD = 0.66 correlation with uncorrected = .99 correlation with multiple choice = .56		CRMS = 0.48 AAD = 0.37 correlation between corrections = 1.00	

Comparison of Classical and Rasch Corrections

Three correction procedures were examined. One, the regression approach was not feasible given the relative emptiness of the data matrix. The remaining two corrections: the linear scaling procedure and the Rasch fair average were employed.

As shown in Table 36, 56.6% of the linearly corrected scores differed from the uncorrected scores by 1 or more points, 10.6% differed by 2 or more points, and 1.2% differed by more than 2 points. By comparison, 51.1% of the Rasch fair average scores differed from the observed average score by 1 or more points, 7.8% differed by 2 or

more points, and 0.5% differed by more than 2 points. The CRMS for the linear correction was 0.84 and the AAD was 0.66; the CRMS for the Rasch correction was 0.88 and the AAD was 0.59. The correlation between the linear corrected score and the multiple choice portion of the examination was .55; the correlation between the Rasch corrected score and the multiple choice portion of the examination was .56.

The two corrections were compared by the same methods as the corrected scores were compared with the uncorrected scores. As shown in Table 36, 74.6% of the linearly corrected scores were identical to the corresponding multifaceted Rasch corrected scores, while 24.6% of the corrected scores differed by 1 point, and only 0.7% differed by 2 points. The CRMS for the comparison was 0.48 and the AAD was 0.37. The correlation between the linear corrected scores and the multifaceted Rasch corrected scores was 1.00. It is clear that the corrections correct in the same direction for a given examinee and that they correct by approximately the same amount.

A Closing Remark

The purpose of this study, as reflected in the two research questions, was the comparison of three approaches in their ability to detect rater variation and to correct for any rater variation found. A strength of the present study was the use of an intact data set with all its natural variation, large size, and sparseness. This data set allowed the three detection approaches and three correction approaches to be compared under the real life conditions in which they would be expected to perform. The approaches are all well established, although the multifaceted Rasch approach is the most recently developed of the three. As indicated in the foregoing, there were some differences and some similarities in results. But given the lack of a simulated data set with known parameters or true results, this study could not identify which, if any approach, was most correct, or which approach was a clear "winner" and which approach a clear "loser". Likewise, this

study did not set out to discover the properties of the variety of new statistics employed. Instead, given nothing more than demonstrations of the superiority of an approach by the proponents of the approach, and the limited research with real data, the intent was to see if the procedures really were different. If the real issue is the fair and equitable treatment of examinees, then the choice of method is moot providing a correction is employed.

CHAPTER VIII CONCLUSIONS

This chapter contains four sections. First, a summary of the study including purpose, data set, and findings is presented. Conclusions of the study are then given. The conclusions section is followed by the limitations of the study. The chapter concludes a discussion of the implications for practice and the implications for research.

Summary of the Study

Purpose and Problems

In spite of the popular assumption found within educational policy debates that performance assessments are inherently superior to traditional standardized multiple choice tests, concerns about rater subjectivity in evaluating student work have not been resolved. The analyses of rater variability have coalesced into three camps: the classical approach, the generalizability approach, and the multifaceted Rasch approach. Three approaches to correction for this variability are: a classical linear scaling, a linear regression, and the use of the multifaceted Rasch model. Little in the way of comparison among the approaches to detection of this variability or among the approaches to correction for this variability has been done. The purpose of this study was to subject a common "real life" data set to the three detection approaches and all three correction approaches to rater variability and then compare the results.

Data Set

The data set consisted of raters' responses to the examinees' written portion of the January sitting of the 1993 Province of Alberta English 33 examination. This data set contained the results of 70 raters who rated the written responses of 4,930 examinees. Each examinee response was scored by 3 different raters applying four, two, and three 5-point scales to Sections I, II, and III of the writing tasks respectively. This marking

scheme produced a data matrix consisting of small cells of 6 examinees by 3 raters by 9 scales. Overall, this scheme produced a data matrix that was 96% empty. The characteristics of this data set meant that while the classical and Rasch analyses were performed on the entire data set the generalizability analysis required the use of a sample rather than the entire data set. A sample of 1067 examinees that included all 70 raters was employed for this purpose.

Analysis and Results

Preliminary analyses. Preliminary analyses were carried out to determine the nature of the nine scales, the feasibility of the proposed generalizability design, and the suitability of the Rasch model for these data. Preliminary analyses included the calculation of means and standard deviations of the nine scales, as well as correlations among the nine scales. Analyses of the generalizability sampling procedure were carried out. Rasch assumptions of unidimensionality, equal discrimination indices, and nonspeededness were tested. Rasch misfit statistics were examined.

Results of preliminary analyses The results of the preliminary analyses supported: an analysis of the data based on the use of nine discrete scales, the feasibility of the aggregation procedure for the generalizability study, and the fit of the data to the Rasch multifaceted model.

Order of analyses. The analyses of rater differences were carried out in the order that they were previously described, that is, the classical approach, the generalizability approach, and the multifaceted Rasch approach. Corrections for rater variability were first made employing a linear scaling that corrects for rater severity and rater central tendency characteristics. A proposed linear regression approach was abandoned as unfeasible. A second correction, in the form of Rasch fair average scores, was examined. Comparisons were made among the results of the three approaches to

detection and between the linear scaling corrected scores and the Rasch fair average scores with the corresponding observed scores.

Classical Analyses and Results

Classical analyses The classical analyses of rater differences consisted of calculation of rater mean severities, rater central tendency deviation, and rater intercorrelations. Raters that differed from their peers were identified for each of the characteristics. Reliability estimates were followed by calculation of standard errors of measurement. The linear scaling correction was carried out at the conclusion of the classical analyses.

Results of classical analyses Raters differed, even after training. Results of the Alexander and Govern (1994) test and subsequent multiple comparisons indicated over 50%, 18 severe and 19 lenient, of the raters differed significantly ($p < .05$) from the mean rater severity. Likewise, the use of the Brown and Forsythe (1974) test and subsequent multiple comparisons indicated over 40%, 16 more central tendency and 13 less central tendency, of the raters differed significantly ($p < .05$) from the mean rater central tendency deviation. Rater intercorrelations indicated a low agreement among raters who marked the same papers. The intercorrelations for all raters was less than .8, with approximately one quarter less than .6; interrater reliability was estimated to be .84. Calculation of coefficient α was .89; the corresponding standard error was 2.08.

Linear scaling correction When the linear scaling was applied 56.6% of the examinees received scores that differed by 1 or more points; 10.6% by 2 or points, and 1.2% by more than 2 points. The correlation of the linearly corrected scores with the multiple choice scores was .55.

Generalizability Analyses and Results

Generalizability analyses. The full generalizability study design was an examinee-by-rater within cell by scales within section. This design was unbalanced due to a different number of scales within sections and the different number of examinees marked by each rater. Consequently, following the suggestion of Shavelson and Webb (1991), a simplified examinees-by-rater-by-scale within bundles (cells) design was considered. These cell variance component results were aggregated to produce mean sample variance component results. The percentage of variance that each facet and interaction accounted for was used to judge the relative importance of each facet. The causes of the large interactions were illustrated using one bundle as an example. Later, standard errors of measurement were calculated from the generalizability coefficients and dependability coefficients. A series of decision studies were carried out.

Generalizability results. The examinee component, 31.8%, and the examinee-by-rater-by-scale, error component, 30.8%, were the largest variance components. The examinee-by-rater and examinee-by-scale components accounted for 11.3% and 17.3% of the variation respectively. The remaining components, rater, scale, and the rater-by-scale interaction, were low accounting for 4.8%, 0.8%, and 3.3% of the total variation respectively. The rater variation, although relatively small, was comparable to the percentage variation (4%) found by Lane and Sabers (1989) and Braun (1988) their studies. The generalizability and dependability coefficients that took into account the variation due to scales and raters, were both .79. In the calculation of the error term for the dependability coefficient, the addition of terms containing the scale variance component and the rater-by-scale variance components resulted in no noticeable change in the value of the coefficient when the dependability coefficient was recorded to two decimal places; this was due to the small magnitudes of these two components. The magnitude of the examinee-rater variance component, 11.3%, a possible indicator of halo

as described by (Popham (1990), indicated that this interaction is an area of rater variation that must be attended to as examinees are not all marked by the same raters.

The value of generalizability coefficient that took into account the variation due to scales was .84. The corresponding standard error of measurement was 2.5. Comparison of this standard error with the standard error from the classical approach revealed that standard error of measurement for the interpretation of examinee scores was approximately 50% larger than a classical approach would predict. If the SEM formula was used when setting the pass / fail cut score, more examinees would be failed if the SEM based on the Hoyt's ANOVA were employed.

It was shown that a reduction in the number of scales leads to a smaller increase in SEM than a comparable reduction in raters.

Multifaceted Rasch Analyses and Results

Multifaceted Rasch analyses The Rasch analysis for raters consisted of the calculation of logit severities, fair average severities, infit and outfit mean square statistics, and Rasch point biserials together with the following group statistics: facet reliability of separation, strata index, and IRR. The Rasch examinee scores, rater scores, and scale scores were examined in relation to their corresponding scores in the observed score metric to better judge the effect of the rater differences.

Multifaceted Rasch results. The multifaceted Rasch analyses compared members of all facets on a common scale. The All Facet Summary revealed that raters were more homogeneous than examinees but less so than scales. The raters differed in Rasch severity as indicated by the fixed chi square test ($\chi^2_{69} = 3792.57$, $p < .001$) and the reliability of separation for raters, .99. Several procedures for determining severity cut off scores were employed. According to the model error analytical procedure, over 80% of the raters (27 severe, 31 lenient) differed from the mean severity; the histogram

procedure (9 severe, 8 lenient) and the natural breaks procedure (9 severe, 7 lenient) both identified under 25% of the raters as different. When the infit t statistic was used to classify raters as to consistency, 34% (24) were judged to be constrained while 30% (21) were judged to be erratic. The Rasch infit mean square guidelines indicated that no rater was too constrained and that 24% (17) of the raters were too erratic. The Rasch point biserials indicated a low but relatively uniform degree of agreement among raters; the values of the mean and standard deviation of the point biserials for raters were identical to the values for scales.

Rasch correction. For the Rasch fair average corrections for examinees, 51.1% of the examinee received scores that differed by 1 or more points; 7.6% differed by 2 or more points, 0.4% differed by more than 2 points. The correlation of the Rasch corrected score with the multiple choice section was .56.

Comparisons

Comparisons of detection approaches. The absolute value for the correlation between the classical rater severity and the rater logit severity was .96. Omnibus tests for rater differences for the classical and for the Rasch agreed that raters differed with respect to severity. In contrast, the variance component for the rater facet in the generalizability study suggested little difference among the raters's severities. Use of the analytical approaches employed following the significance of the classical and Rasch omnibus tests resulted in the identification of a large but unequal percentages of severe and lenient raters (25.7% severe, 27.1% lenient versus 38.6% severe, 44.2% lenient). Working with the 10 most severe and the 10 most lenient raters revealed greater agreement: 9 out of 10 raters were identified by both procedures for both severe and lenient rater groups.

Rater central tendency deviations (CTDs) correlated .76 with Rasch infit statistics. There is no measure of central tendency within generalizability theory. The omnibus test

for the classical approach indicated raters differed. There was no omnibus test for the Rasch approach. Use of the analytical approaches employed in the classical and Rasch approaches resulted in the identification of a large but unequal percentages of over consistent and erratic raters (22.9% constrained, 18.6% erratic versus 34.3% constrained, 30.0% erratic). Seven out of the 10 most constrained raters identified by the classical approach were identified by the Rasch approach and 5 out of the 10 least consistent raters identified by the classical approach were identified by the Rasch approach as erratic. The moderately large examinee-rater interaction variance component might be an indication of erratic raters but this comparison at the examinee-rater level was not made in this study.

Scale test-score correlations and Rasch scale point biserials correlated .96, but differed in both magnitude and dispersion while classical rater intercorrelations and Rasch point biserials correlated only .31. Given the lack of relationship between these variables, no further comparisons were made.

The two correction procedures produced very similar results. Comparison with the observed scores revealed that slightly more than 50% of the examinees received a corrected mark that differed by 1 or more points with either correction, 10.6% and 7.6% of the examinees received a mark that differed by 2 or more points with the classical and Rasch corrections, respectively. Approximately 1% or less had corrected marks that differed by more than 2 points with either correction.

Conclusions

The system of examinees, raters, and scales analyzed in this study proved to be typical of other rating systems described in the literature. All three approaches to detection of rater differences yielded results which were similar to results found in other studies in which the approaches considered in the present study were used. More

specifically, raters differed from one another on a number of characteristics: they displayed differing severities, they displayed differing degrees of central tendency error, and they have low measures of agreement with each other. When corrections were applied, both the classical linear equating and the Rasch fair average produced corrections of one or more points for a majority of the population.

As a way of judging the approaches employed in this study, two questions were formulated in Chapter I. The first question addressed rater variation while the second addressed corrections for that variation. The questions and related results follow.

The First Question

The first question raised in Chapter I was,

Do Generalizability theory and a multifaceted Rasch analysis produce importantly different indications of rater variability than the indications provided by the classical approach when these analyses are applied to a typical data set produced by students being assessed on a performance assessment task administered on a province-wide basis?

In the first question "importantly different" referred to results that would yield different conclusions to be drawn based on the results yielded by the analysis used. The CTT measures and the MFRM measures both agreed that raters as a group differed but disagreed on how many raters differed. Several Rasch indices: reliability of separation, the strata index, and IRR all appear to provide inflated measures when large numbers of observation per facet member are employed. In contrast, the generalizability theory approach indicated raters did not vary considerably with respect to severity but instead identified the examinee-rater interaction (halo) as a major source of total variation.

A second characteristic of an individual rater was the tendency to employ marks near to the rater's mean mark. This statistic used for this central tendency deviation, the CTD, indicated that some raters awarded relatively few marks different from those raters' respective mean marks awarded while others were more willing to use a wider range of

marks. The MFRM approach was different from the classical approach in that it focused on the identification of raters that were both too consistent and too inconsistent in rating practices. CTD and infit were compared as some similarity was expected. The moderately high correlation suggested that CTD and infit were not the same characteristic but did share substantial common variance. However, the extent of relationship was not as strong as it was for rater severity. Extreme raters identified by one index tended to be identified as extreme by the other index.

The third broadly defined characteristic considered was the agreement among raters who marked the same papers. The CTT approach and Rasch approaches indicated agreement was low to moderate. These two approaches differed more on indicators of agreement than for the other two characteristics. This form of rater agreement is not addressed by the generalizability approach.

In comparison to generalizability estimates, traditional CTT estimates of reliability are incapable of simultaneously identifying multiple sources and magnitudes of error. Consequently they yield overestimates of reliability and correspondingly underestimates of the standard error of measurement. GT is capable of providing a variety of SEMs to suit a variety of measurement situations. GT is clearly superior to CTT in this aspect. The MFRM can place all facets on a common scale for comparison. MFRM provides a variety of reliability indices but their interpretation through use has not been developed. One possible reason for the lack of use of several of these indices is the redundancy among them due to the mathematical relationships among them (see pp. 60-65). Both GT and MFRM offer some clear advantages that cannot be obtained through the use of a CTT approach alone.

In summary, the practitioner cannot assume that all or any of the three approaches will produce similar results. While some results were similar across approaches, others were different. The approach or combination of approaches taken to assess rater variation

and to estimate reliability and SEM must be carefully chosen with consideration of the type of information required.

The Second Question

The second question raised in Chapter I was,

Do either a linear scaling, a linear regression, or a multifaceted Rasch analysis produce importantly different indications of examinee ability than the those of the uncorrected classical approach when these analyses are applied to a typical data set produced by students assessed on a performance assessment task administered on a province-wide basis?

The linear regression approach was not carried out due the nature of the relatively empty matrix is such that although the procedure is known, it would require a larger amount of time and computer RAM to complete this analysis.

In both cases of correction, "importantly different" was defined as results that differed by a one or more points in a total score of forty five marks. Both the linear scaling approach and the Rasch approach yielded a mark that differed by one or more points for more than 50% of the examinees. Thus half the students received corrections that were importantly different when either the linear scaling or the multifaceted Rasch model corrections were applied. The two corrections yielded results that were very similar, 75% of the examinees received the same corrected score by both procedures. The correlation between the corrected scores was 1.00. Practitioner's should apply a correction to rated data to correct for differences in rater characteristics that were shown to be present.

Limitations

The use of the large scale sparse data set lead to several limitations in this study that were due to the size of data matrix and its relative emptiness. The most severe limitation were the inability to employ a full generalizability model and complete a regression correction. Even with the generalizability approach taken, the sheer number of

complete cells is such that examination of interaction effects, that is, rater-examinee, rater-scale, and examinee-scale, cannot reasonably be carried out.

This study was carried out on one sitting of one examination of one subject. It is not known from this study whether similar results would be found under a replication with the same subject at a different sitting, with a similar subject like social studies, or with a dissimilar subject such as mathematics.

Implications

Implications for Practice

There are a number of implications for practice. These will be discussed in the order that the analyses were conducted. First of all, use of Cronbach's α , in which rater variation is ignored, results in an overestimation of the actual reliability of the scores. If the classical analysis continues to be the only approach taken, the Spearman-Brown prophecy correction should be employed in which the mean rater intercorrelation is taken as the reliability of a single rater (Winer, 1971, p. 289).

Generalizability theory was demonstrated to have unique strengths in relation to the other two approaches to analysis. The large examinee-rater interaction effect was not detected by other methods; raters who are idiosyncratic in their markings of a variety of examinee papers should be identified in order to ensure the fair marking of all papers. Generalizability theory offers the opportunity to analyze rating systems in a manner that no other approach can. The present disadvantage to the practitioner is a lack of a suitable program for the analysis of large, sparse data sets.

The multifaceted Rasch model was demonstrated to be suitable for analyses involving large sparse data sets. However, some MFRM statistics, in particular, rater infit and rater Rasch Point biserial, require clarification before practitioners can routinely

interpret them. Another practical downside was the very lengthy computer running time. One analysis of this data set on a micro computer with a 50 MHz clock speed required approximately 8 hours.

Lastly, both the classical linear equated scores and the Rasch fair average were importantly different from the uncorrected scores and these two corrections produced very similar results for the individual examinees. Both procedures account for rater severity and consistency. Consequently, in the interest of fairness and equality to examinees, these corrections should be routinely employed, with the selection of correction procedure dependent on the approach taken.

Implications for Future Research

Some implications for research relate to the findings of the analyses themselves; other implications are result of the comparisons between approaches. Implications will be considered in the order just described.

The results of the various approaches of this study should be investigated for other similar subjects, that is, other humanities subjects and for subjects in the science and mathematics areas. Additionally, facets that were not analyzed in this study such as gender of examinee, gender of rater, ethnicity, and occasion should be analyzed. Differences within these facets would present serious problems as these are situations in which differences should not occur, and if they do occur.

The large examinee-rater interaction effect noted in the generalizability results, should be explored in order to first identify raters who are prone to large interaction effects, and then to investigate the causes of the interaction. The research should consider the characteristics of raters or alternatively characteristics of the examinee. It is found that the examinee-by-rater-by-scale, error variance component was large. The sources of variation within that component should be investigated.

The interpretation of the Rasch fit statistics requires clarification, particularly in relation to halo, as this study found the rater fit statistics closely correlated with rater CTD. Likewise, further research is needed to link Rasch reliability like statistics such as rater point biserial, rater separation, and IRR, to classical definitions.

Simulation studies should also be carried out to establish which correction, linear scaling or multifaceted Rasch, can best reproduce the original data from a matrix in which much of the data has been deleted. Now that the linear scaling corrections and Rasch corrections are known to be of the same magnitude, a simulation study would allow the test of which of the two corrections gives the more accurate correction for those examinees that did not receive the same corrected scores by the two methods.

REFERENCES

- Alberta Education. (1992). 1992-93 English 33 Bulletin Update: Diploma Examinations Program. Edmonton, AB: Author
- Alberta Education. (1994). 1994-95 English 33 Information Bulletin: Diploma Examinations Program. Edmonton, AB: Author
- Alberta Education. (1993). January 1993 Marker's Manual: English 33 Diploma Examinations Program. Edmonton, AB: Author
- Alexander, R. A., & Govern D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. Journal of Educational Statistics, 19, 91-101.
- Allen, M. J., & Yen W. M. (1986). Introduction to measurement theory. Monterey CA: Brooks/Cole Publishing.
- Alliger, G. M., & Williams, K. J. (1992). Relating the internal consistency of scales to rater response tendencies. Educational and Psychological Measurement, 52, 337-343.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.
- Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.
- Aschbacher, P. R. (1991). Facing the challenges of a new era of educational measurement. Applied Measurement in Education, 4, 275-288.
- Baxter, G. P., Shavelson, R.J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1-17.
- Becker, M., Hess R. K., & Gibney, V. (1993, April). Large-scale assessment in writing: Factors influencing scaling of writer's performance. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley.
- Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, 22, 41-52.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.
- BMDP Statistical Software Inc., (1990). BMDP Statistical Software Manual Volume 2. [Computer program manual]. Los Angeles: Author.

- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277-289.
- Brooks, T. E., & Twing, J. S. (1992, April). A many-faceted approach to the Assessment of math performance data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA
- Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of American Statistical Association, 69, 364-367.
- Candell, G. L., & Ercikan, K. (1992, April). Assessing the reliability of the Maryland School Performance Assessment Program using generalizability theory. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.
- Cason, G. J., & Cason, C. L. (1984). A deterministic theory of clinical performance rating. Evaluation and the health professions, 14, 221-247.
- Chauvin, S. W., Ellett, C. D., Loup, K. S., Naik, N. S. (1992, April). The effects of assessment demand characteristics on the generalizability of classroom-based assessments of teaching and learning: Will the real reliability coefficient please stand up?. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.
- Coffman, W. E., (1971). Essay examinations. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 271-302). Washington, DC: American Council on Education.
- Crocker, L., & Algina J. (1986). Introduction to classical and modern test theory. New York: Holt, Reinhard and Winston.
- Cronbach, L. J., Gleser, F. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.
- Cronbach, L. J., Rajaratnam, N. & Gleser, B. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.
- Crowley, S. L., Thompson, B. & Worchel, F. (1994). The Children's Depression Inventory: a comparison of generalizability and classical test theory analyses. Educational and Psychological Measurement, 54, 705-713.
- de Grujter, D. N. M. (1984) Two simple models for rater effects. Applied Psychological Measurement, 8, 213-218.

- Darling-Hamond, L. (1994). Symposium: Equity in educational assessment. Harvard Educational Review, 64, 5-30.
- Donoghue, J. R. (1994). An empirical examination of the IRT information of polychotomously scored reading items under the generalized partial credit model. Journal of Educational Measurement, 31, 291-311.
- Elder, R. W. (1991). Application of generalizability theory to the analysis of designs involving random nested dependent variables. Unpublished masters thesis, University of Alberta.
- Engelhard, G. Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.
- Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, 31, 93-112.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn. (Ed.), Educational measurement (3rd ed.). (pp. 105-146). New York: Macmillan.
- Ferrara, S. (1993, April). Generalizability theory and scaling: Their roles in writing assessment and implication for performance in other content areas. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA
- Fischer, A. G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. The American Journal of Occupational Therapy, 47, 319-329.
- Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 237-270). Washington, DC: American Council on Education.
- Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48, 157-184.
- Glass, G. V., & Hopkins, K. D. (1984). Statistical methods in education and psychology. (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.
- Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.
- Guilford, J. P. (1954). Psychological methods. New York: McGraw-Hill.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn. (Ed.), Educational measurement (3rd ed.). (pp. 13- 104). New York: Macmillan.
- Hambleton, R. K. (1987). The three-parameter logistic model. In D. L. McArthur. (Ed.), Alternative approaches to assessment of achievement. (pp. 129- 158). Norwell MA: Kluwer Academic Publishers.

- Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory. (pp. 71-94). Vancouver BC: Educational Research Institute of British Columbia.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA Sage Publications.
- Hambleton, R. K., & Wright, B. D. (1992, April). Item response theory in the 1990s: Which models work best? Invited debate at the annual meeting of the National Council of Measurement in Education, San Francisco CA
- Hess R. K. & Olsen, R. (1993, April). Performance based assessment in writing: Detecting bias in raters and prompts. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA
- Houston, W. M, Raymond, M. R, & Svec, J. C. (1991) Adjustments for rater effects in performance assessment. Applied Psychological Measurement, 15, 409-421.
- Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.
- Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. Journal of Educational Measurement, 21, 277-281.
- Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research, 60, p. 237-263.
- Huynh, H. & Ferrara, S. (1994). A comparison of equal percentile and partial credit equating for performance-based assessments composed of free-response items. Journal of Educational Measurement, 31, 125-142.
- Johnson, S. (1989). National assessment: The APU science approach. London: Her Majesty's Stationary Office.
- Julian, M. W. & Searcy, C. A. (1995, April). Regression procedures for the detection of rater effects in direct writing assessment. Paper presented at the annual meeting of American Educational Research Association, San Francisco, CA
- Kenyon, D. M., & Stansfield, C. W. (1992, April). Examining the validity of a scale used in a performance assessment from many angles using the many-faceted Rasch model Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.
- Kirk, R. E. (1990). Statistics: An introduction. (3rd ed.). Fort Worth TX: Holt, Rinehart and Winston.
- Kromrey, J., Bullock, D., Chason, W., Du Bose, P., Harrison, B. (1992, April). Assessment of reliability of holistic scoring of foreign language and speech performance examinations. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.
- Koch, W.R. (1983). Likert scaling using the graded response latent trait model. Applied Psychological Measurement, 7, 15-32.

- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational measurement: Issues and Practice, 13, (3) p.5 - 16.
- Lane, S. & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. Applied Measurement in Education, 2, 195-205.
- Littefield, J. H., Harrington, J. T., Anthracite, N. E., & Garman, (1981). A description and four-year analysis of a clinical clerkship evaluation system. Journal of Medical Education, 56, 334-340.
- Linacre, J. M. (1987). An extension of the Rasch model to multi-faceted situations. Chicago: University of Chicago, Department of Education..
- Linacre, J. M. (1989). Many-facet Rasch measurement Chicago: MESA Press.
- Linacre, J. M. (1992). Facets: version 2.62 [computer program]. Chicago: MESA Press.
- Linacre, J. M. (1992). Facform: version 1.22 [computer program]. Chicago: MESA Press.
- Linacre, J. M. & Wright, B. D. (1992). A user's guide to Facform: Data formatting computer program for Facets. Chicago: MESA Press.
- Linacre, J. M. & Wright, B. D. (1992). A user's guide to Facets: Rasch measurement computer program. Chicago: MESA Press.
- Linn, R. L. (Ed.). (1989). Educational measurement (3rd ed.). New York: Macmillan.
- Linn, R. L., & Burton, E. (1994). Performance - assessment: Implications of task specificity. Educational measurement: Issues and Practice, 13, (1) p.5 - 15
- Linquist, E. F. (Ed.). (1951). Educational measurement. Washington, DC: American Council on Education.
- Lord, F. M. (1952). A theory of test scores. (Psychometric Monograph No. 7). Psychometric Society.
- Lunz, M. E. (1993, April). Grading Plans and Judges. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA.
- Lunz, M. E. & Stahl, J. A. (1990). Judge consistency and severity across grading periods. Evaluation and the health professions, 13, 425-444.
- Lunz, M. E. & Stahl, J. A., (1992). New ways of thinking about reliability. Professions Education Researcher Quarterly, 13, (4). 16-18
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of Judge severity on examination scores. Applied Measurement in Education, 3, 331-345.
- Marcoulides, G. & Linacre, J. M. (1993, April). Large-scale assessment in writing: Factors influencing scaling of writer's performance. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA

- Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.
- Meredith, V. H., & Williams, P. L. (1984). Alternative assessment in a high-stakes environment. Educational measurement: Issues and Practice, 3, (1) p.11-15
- Michael, W. B., Cooper, T., Shaffer, P., & Wallis, E. (1980). A comparison of the reliability and validity of ratings of student performance on essay examinations by professors of English and by professors of other disciplines. Educational and Psychological Measurement, 40, 183-195.
- Microsoft Corporation., (1992). Microsoft excel user's guide 1. [Computer program manual]. USA: Author.
- Microsoft Corporation., (1992). Microsoft excel user's guide 2. [Computer program manual]. USA: Author.
- Microsoft Corporation., (1992). Microsoft excel function reference. [Computer program manual]. USA: Author.
- Miller, M. D. & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. Educational measurement: Issues and Practice, 12, (2) p.9-15
- Mullis, I. V. S. (1984). Alternative assessment in a high-stakes environment. Educational measurement: Issues and Practice, 3, (1) p.9-15
- Muraki, E. (1992) A generalized partial credit model. Applied Psychological Measurement, 16, 159-176.
- Myford, C. M. (1992, April). Identifying, representing, and controlling for judge differences in performance assessments: An example from the visual arts. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA
- Nandakumar, R. (1994). Assessing dimensionality in set of item responses - Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.
- Nyberg, A. M. (1987). A reliability study of high school essay scoring. Unpublished doctoral dissertation, University of Alberta, Edmonton.
- O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. Psychological Bulletin, 89, 570-574.
- Popham, W. J. (1990). Modern educational measurement: a practitioners perspective (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.
- Principles for fair student assessment practices for education in Canada . (1993)
Edmonton AB: Joint Advisory Committee.
- Ramsey, P. H. (1994). Testing variances in psychological and educational research. Journal of Educational Statistics, 19, 23-42.
- Rasch, G. (1960 / 1972). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

- Rasch, G. (1961) On general laws and the meaning of measurement in psychology. In J. Neyman. (Ed.), Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability, Berkeley: University of California Press.
- Raymond, M. R. & Houston, W. M. (1990). Detecting and correcting for rater effects in performance assessment. (Report No. ACT-RR-90-14). American College Testing Program
- Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measurement, 47, 13-26.
- Raymond, M. R. & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. Journal of Educational Measurement, 30, 253-268.
- Raymond, M.R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. Evaluation and the health professions, 14, 100-122.
- Reardon, S.F., Scott, K., & Verre, J. (1994). Symposium: Equity in educational assessment. Harvard Educational Review, 64, 1-4.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.
- Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R.J. (1993). On the stability of performance assessments. Journal of Educational Measurement, 30, 41-53.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 34, (2, Whole No. 17).
- Samejima, F. (1973). Homogeneous case of the continuous response model. Psychometrika, 38, 203-219.
- Sax, G. (1989). Principles of educational and psychological measurement and evaluation. (3rd ed.) Belmont: Wadsworth.
- Schultz E. M. & Linacre, J. M. (1995, April). A comparison of the many-facet Rasch model and generalizability procedures for estimating variance and reliability in writing assessment. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA
- Shavelson, R.J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. Applied Measurement in Education, 4, 347-362.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.
- Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA Sage Publications.
- Smith, P. L. (1978). Sampling errors of variance components in small multifacet generalizability studies. Journal of Educational Statistics, 3, 319-346.

- Smith, P. L., & Leucht, R. M. (1992). Correlated effects in generalizability studies. Applied Psychological Measurement, 16, 229-235.
- Smith, R. M., Schumacker, R. E., & Bush, M. J. (1995, April). Using Item Mean Squares to Evaluate Fit to the Rasch Model. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.
- SPSS Inc. (1990). SPSS Reference Guide. [Computer program manual]. Chicago: Author
- Stahl, J. A., & Lunz, M. E. (1993, April). A comparison of generalizability theory and multi-faceted Rasch measurement. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA
- Stalnaker, J. M. (1951). The essay type of examination. In E. F. Lindquist (Ed.), Educational measurement (pp. 495-530). Washington, DC: American Council on Education.
- Stiggins, R. J. (1991). Facing the challenges of a new era of educational measurement. Applied Measurement in Education, 4, 263-274.
- Tatum, D. S. (1992, April). Controlling for judge differences in the measurement of public speaking ability. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA
- Thorndike, R. L. (Ed.). (1971). Educational measurement. (2nd ed.). Washington, DC: American Council on Education.
- Twing, J. S. & Williams, K. T. (1992, April). An investigation of writing assessment using a many-faceted Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA
- Welch, C. J. & Miller, T. R. (1995, April). Scaling holistically scored writing assessments using a many-faceted Rasch model. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA
- Winer, B. J. (1971). Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill
- Wilson, H. G. (1988). Parameter estimation for peer grading under incomplete design. Educational and Psychological Measurement, 48, 69-81.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370 (Available from the Rasch Measurement SIG, 5835 S. Kimbark Avenue, Chicago IL 60637-1609).
- Wright, B. D., & Linacre, J. M. (1992). A user's guide to BIGSTEPS: Rasch-model computer program. Chicago: MESA Press.
- Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

Du, Y. (1995, April). Measuring student writing abilities in a large-scale writing assessment. Paper presented at the Eighth International Objective Measurement Workshop, Berkeley, CA

Zegers, F. E. (1991). Coefficients for interrater agreement. Applied Psychological Measurement, 15, 321-333.

Appendix A

January 1993 Marker's Manual: English 33 Diploma Examinations Program

RESPONSIBILITIES OF GROUP LEADERS

it is each Group Leader's responsibility to

1. assist Student Evaluation Branch staff in training markers. This will include the following:
 - review training papers and the rationale for scores in order to clarify for markers the relationship between the scoring guides and the students' work
 - chair the small-group discussions following the scoring of training papers
 - support new markers in their orientation to marking procedures
2. Answer questions and concerns raised by markers in your group and refer to your section supervisor questions or concerns that require further clarification
3. Chair small groups during reliability reviews
4. Assist Student Evaluation Branch staff with problem papers and special cases
5. Ensure that papers considered to be Insufficient are tagged appropriately and are referred to supervisors for confirmation
6. Score papers requiring a fourth reading
7. Notify the supervisor who is assigned to your group of any circumstances that will affect a marker's attendance

RESPONSIBILITIES OF MARKERS

As a marker, it is your responsibility to

1. Apply the scoring guides impartially, independently, and consistently to all papers (including specially processed papers such as large print papers, papers translated into Braille, word processed papers, and papers written with the assistance of a scribe)
2. Complete the scoring sheets as instructed
3. Place your marker number in the appropriate box on the back cover of the examination booklet
4. Refrain from marking a paper if personal biases (i.e., handwriting, political or religious reference) might interfere with your impartial judgment of the student's writing
5. Refer to your group leader any paper of special concern, for example, any paper in which prepared materials have been included or any paper in which departures from regulations are suspected
6. Refer papers that you consider to be Insufficient to your supervisor for confirmation
7. Refrain from discussing students' written work during independent marking and at such times as discussion may prove prejudicial to a successive marking
8. Participate professionally in reliability review sessions and general discussions as scheduled
9. Inform your group leader of any circumstance that will require you to be late or absent from a marking session
10. Provide all required information necessary to the processing of your expense claim before you leave the marking session (see time sheet page 43)
11. Adhere to the policy regarding children in need of protective services (see papers 11-13)

PAPER FLOW

1. The papers have been divided into bundles of six.
2. You will pick up a bundle of unmarked papers from the distribution centre and return to your marking station.
3. **BE SURE THAT THE INDEX NUMBER ON THE BACK OF EACH EXAM BOOKLET CORRESPONDS WITH THE INDEX NUMBER ON THE SCORING SHEET FOR THAT EXAM BOOKLET.** Once you have completed all sections of the scoring sheet (see page 15), remove that sheet from the booklet and set it aside. Leave the other scoring sheets in the booklet.
4. Enter your marker number in the appropriate box on the back of each exam booklet.
5. When you have completed the marking of all six papers in the bundle, check your scoring sheets to make sure that each has been completed properly.
6. Tag and refer to your section supervisor papers that you deem to merit a score of **INSUFFICIENT**.
7. Place the completed scoring sheets on top of the bundle and carefully replace the elastic bands around the bundle.
8. Return your bundle of marked papers to the distribution centre and pick up another bundle of unmarked papers.

INSTRUCTIONS FOR DEALING WITH INSUFFICIENT CATEGORY

There will be very few INS papers. Almost ALL students BELIEVE that they are doing the assignment. The descriptors for (1) POOR will fit almost all weak or generally misguided work. An INS for Thought and Detail requires an INS on ALL other scales.

Regarding INS for length: Keep in mind that a very brief but clear statement may well indicate that the student comprehends the task; however, remember that the assignments require developed ideas. If there is no development, it is not likely that a fair assessment of the student's performance on each scale can be made.

All cases are different. Judgments must be made in accordance with the individual situation.

PROCEDURES: A response that the marker considers Insufficient can be discussed with the group leader if he or she is available. All insufficient scores must be confirmed by a supervisor. Tag the paper in question indicating

- Index number
- Assignment
- Reason for INS
- Marker #, Table #
- Date

and leave the paper with a supervisor. Paper will be returned for discussion and tags removed.

RELIABILITY REVIEW: PURPOSE AND PROCEDURE

PURPOSE: To promote inter-marker reliability and consistent application of the scoring guides.

Reliability reviews provide markers with the opportunity to compare individual scores and rationales with those of others, and to review the scoring criteria and their application. The discussions during reliability reviews should serve to remind markers that *the consistent application of the scoring descriptors to all papers is of utmost importance.*

The main purpose of a reliability review is to ensure that a paper exhibiting the characteristics of a particular scoring descriptor will be awarded the mark associated with that scoring descriptor by all markers.

For example, a paper having features in a given scoring category (e.g. Thought and Detail) that clearly fit the descriptor for Proficient should be awarded a 4 by all markers. Examples will arise, however, of papers that appear to exhibit the characteristics of two scoring descriptors. In such cases, markers are instructed to use their judgement and to apply the scoring descriptor that BEST describes the features of the paper being scored.

PROCEDURE

1. Each person in the reliability review group reads and scores the preselected paper independently.
2. The group leader records the initial marks assigned by each marker. The tally of these initial marks will be posted to enable each marker to compare his or her scores with those of the entire marking group.
3. Following the recording of the initial marks, the group leader opens a discussion about the scoring of the reliability review paper, first inviting the marker awarding the highest mark to defend his/her decision in terms of the scoring guides and the student's work. A marker awarding a discrepant mark provides the second defence. The discussion should focus on each of the scoring categories (Thought and Detail, Organization, etc.), and the features of the student's work that illustrate the scoring descriptors.
 - A. When marks of all the markers in the group are in agreement for any one category, members should determine the source of agreement by matching components of the paper to the descriptors of the scoring guides.
 - B. When marks prove to be divergent or discrepant* for a category, the marker who has assigned the divergent or discrepant mark should point out the characteristics of the student's work that prompted the marking of this paper in this manner. The other markers may then contribute to the discussion of the paper with respect to the category in question. *Consensus should be reached based on the scoring descriptors as they apply to the student's work.*

- C. Throughout the discussion, markers are requested to respond to the paper *as if the student writer were present*. Each marker should ensure that all references relate to the paper at hand and to the scoring guides. References to content of other papers read during independent marking must be avoided. Comparison to student compositions written in dissimilar situations must be avoided.
4. The group leader records the marks assigned by each marker after the discussion and forwards the summary of initial and post-discussion marks to the supervisor. Results of the post-discussion marks will be tallied and posted with the initial results.
- * A discrepant mark differs from the median mark by more than one point. For example, if a paper receives marks of 3, 3, and 1 in a scoring category, the median mark is 3 and the mark of 1 is discrepant. A divergent mark differs from the median by only one point. For example, if a paper receives marks of 3, 4, and 4 in a scoring category, the median mark is 4 and the mark of 3 is divergent.

INDIVIDUAL MARKER SUMMARY SHEETS

To assist markers in scoring papers as reliably as possible, individual marker summary sheets are distributed to markers periodically throughout the marking session. A final summary will be mailed to all markers for their personal interest after the marking session is completed. These summary sheets are designed to provide information about the scores assigned by each individual marker and to allow the marker to compare them to scores assigned to all other papers and to papers scored in common.

"Number of Papers Scored" (see ① on sample Individual Marker Summary Sheet) is calculated from a cut-off point including only papers that have had three readings. *This figure will not match the marker's current total of papers scored.*

Individual marker statistics are not completely comparable with those of other markers because each marker does not score the same papers in the same combination with other markers. Therefore, considerable variation among markers is expected and normal (see ② on sample Individual Marker Summary Sheet).

Inserts ③ and ④ on the sample Individual Marker Summary Sheet provide detailed explanation of scoring frequency on all papers scored ③, and marker variation on papers scored in common ④.

Caution should be exercised in drawing conclusions on the basis of these statistics alone. Rather, these summary reports are intended to alert markers to discrepancies that may exist between the scores they have awarded and the scores awarded by others who have marked the same papers.

In combination with individual scores awarded on reliability reviews compared with the whole group, these statistics can draw attention to scoring practices that are potentially discrepant.

SAMPLE INDIVIDUAL MARKER SUMMARY SHEET

INDIVIDUAL MARKER DAILY SUMMARY SHEET, ENGLISH 35

DATE: 28/01/93

MARKER: 000

Number of Papers Scored: 20 Average for All Markers: 29.

Percent of All Papers Requiring Rescoring: 1.0%

The following table will allow you to compare the scores you have given to the scores other markers have given for each category for each scale, and how often you awarded scores in each category for each scale, and how often other markers awarded scores in each category for each scale. Since a different selection of papers was marked by each marker, some variation is expected.

NOTE: Zero scores include blanks, No Response, and Insufficient rankings.

Scale By Scale Category Frequencies In Percent

Scale Name	0	1	2	3	4	5
Personal Response Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Choice	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Convention	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Functional Writing Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Response To Visual Command Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Personal Response Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Choice	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Convention	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Functional Writing Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Response To Visual Command Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000

Personal Response Assignment: Thought and Detail Scale, 17.2% of all scores awarded were 2's, 50.9% were 3's, and 25.2% were 4's. Marker 000 had more 2's (25.0%), more 3's (64.3%), and fewer 4's (10.7%) than did all other markers.

Personal Response Assignment: Thought and Detail Scale, other markers gave the median score 72.5% of the time. They gave scores 1 below the median 12.4% of the time, and scores 1 above the median 13.0% of the time. Marker 000 was on the median 78.6% of the time, and 1 below the median 21.4% of the time, and 1 above the median 0% of the time. Scores + or - one from the median do not contribute to the number of papers requiring rescoring. However, scores of + or - two from the median contribute to the number of papers requiring rescoring if such discrepancies occur on several scales.

The second table indicates the amount and direction of variation from the median (middle) score of the three scales in each category. In this table, your scores are compared to scores awarded by other markers to the same papers. A negative variation is a score below the median.

Scale By Scale Frequencies Of Marker Variation From Median In Percent

SCALE NAME	-5	-4	-3	-2	-1	0	1	2	3	4	5
Personal Response Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Choice	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Matters of Convention	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Functional Writing Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Response To Visual Command Thought and Detail	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Organization	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000
Writing Skills	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000	0000000000

Section I: Personal Response to Literature - Scoring Guide

It is important to recognize that student responses to the Personal Response Assignment will vary from writing that treats personal views and ideas analytically and rather formally to writing that explores ideas experimentally and informally. Consequently, evaluation of the personal response on the diploma examination will be in the context of Louise Rosenblatt's suggestion.:

The evaluation of the answer would be in terms of the amount of evidence that the [student] has actually read something and thought about it, not a question of whether necessarily he has thought about it in the way an adult would, or given an adult's "correct" answer.¹

Thought and Detail (curriculum concepts 1, 3, 4, 6, 7, 8, 9, 12)

When marking Thought and Detail, the marker should consider how effectively

- the assignment is addressed
- the detail supports and/or clarifies the response

- 5 **EXCELLENT:** An insightful understanding of the reading selection(s) is effectively demonstrated. The student's opinion, whether directly stated or implied, is perceptive and is appropriately supported by specific details. Support is precise and thoughtfully selected.
 - 4 **PROFICIENT:** A well-considered understanding of the reading selection(s) is appropriately demonstrated. The student's opinion, whether directly stated or implied, is thoughtful and is supported by details. Support is well defined and appropriate.
 - 3 **SATISFACTORY:** A defensible understanding of the reading selection(s) is clearly demonstrated. The student's opinion, whether directly stated or implied, is conventional but is plausibly supported. Support is general but functional.
 - 2 **LIMITED:** A vague understanding of the reading selection(s) is evident but is not always defensible or sustained. The student's opinion may be superficial, and support is scant and/or vague, and/or redundant.
 - 1 **POOR:** An implausible conjecture concerning the reading selection(s) is suggested. The student's opinion, if present, is irrelevant or incomprehensible. Support is inappropriate, inadequate, or absent.
- INS INSUFFICIENT:** The marker can discern no evidence of an attempt to fulfil the assignment, or the writing is so deficient in length that it is not possible to assess thought and detail.

¹ Rosenblatt, Louise. "The Reader's Contribution in the Literary Experience." An interview with Lionel Wilson in *The English Quarterly* 1 (Spring, 1981): 3-12.

Section I: Personal Response to Literature - Scoring Guide (continued)

Organization (curriculum concepts 2, 3, 4)

When marking Organization, the marker should consider how effectively the writing demonstrates

- unified and consistent development
- clear and coherent order

- 5 **EXCELLENT:** The beginning is constructed to provide direction for the reader and/or to encourage further reading. The ideas and situations are developed by sentences and paragraphs that flow smoothly and coherently to an appropriate and effective conclusion.
- 4 **PROFICIENT:** The beginning is constructed to provide direction for the reader. The ideas and situations are developed by sentences and paragraphs that are coherently related. The conclusion is appropriate.
- 3 **SATISFACTORY:** The beginning, development, and conclusion are functional. Sentences and paragraphs are generally related, but coherence may falter on occasion.
- 2 **LIMITED:** The beginning and/or conclusion are nonfunctional. Relationships between sentences and between paragraphs are frequently unclear.
- 1 **POOR:** The beginning is vague and/or unfocused. The conclusion, if present, is vague and/or unfocused. Sentences and paragraphs are not coherently related.

Section I: Personal Response to Literature - Scoring Guide (continued)

Matters of Choice (curriculum concepts 3, 4)

When marking Matters of Choice, the marker should consider the extent to which the writing demonstrates effectiveness of

- diction, including connotative language, imagery, idiomatic expressions, dialect
- syntax, including such choices as parallelism, balance, inversion, sentence length and variety

- 5 **EXCELLENT:** Diction is appropriate and precise. Many sentences have been successfully structured for effect. Choices evident in the writing are usually effective and sometimes polished.
- 4 **PROFICIENT:** Diction is appropriate and generally effective. Many sentences appear to have been purposefully structured for effect. Choices evident in the writing are often effective.
- 3 **SATISFACTORY:** Diction is appropriate but may be general rather than specific. Sentence structure is generally straightforward and clear. Choices evident in the writing are occasionally effective.
- 2 **LIMITED:** Diction is inaccurate and/or vague. Sentence structures are misused to such an extent that clarity suffers. Choices evident in the writing are usually ineffective.
- 1 **POOR:** Diction is inaccurate and/or vague. Sentence structures are misused to such an extent that clarity suffers. Choices evident in the writing are usually ineffective.

Section I: Personal Response to Literature - Scoring Guide (continued)

Matters of Convention (curriculum concepts 3, 4)

When marking Matters of Convention, the marker should examine the writing for correctness of

- mechanics (spelling, punctuation, capitalization, indentation, etc.)
- grammar (agreement of subject-verb/pronoun-antecedent, pronoun reference, etc.)

PROPORTION OF ERROR TO COMPLEXITY AND LENGTH OF RESPONSE MUST ALSO BE CONSIDERED.

- 5 EXCELLENT: This writing is essentially free from errors in mechanics and grammar. Errors that may be present do not reduce the clarity of communication.
- 4 PROFICIENT: This writing is essentially free from errors in mechanics and grammar. Seldom do any errors that may be present reduce the clarity of communication.
- 3 SATISFACTORY: This writing has occasional errors in mechanics and grammar. A few of these errors may reduce the clarity of communication.
- 2 LIMITED: This writing has frequent errors in mechanics and grammar. Many of these errors reduce the clarity of communication.
- 1 POOR: This writing has numerous errors in mechanics and grammar that are both noticeable and jarring. Most of these errors severely reduce the clarity of communication.

Section II: Functional Writing - Scoring Guide

Thought and Detail (curriculum concepts 1, 3, 4, 5)

When marking Thought and Detail, the marker should consider

- how well the assignment is addressed and the purpose fulfilled
- awareness of audience (appropriateness of tone)
- effectiveness of development

- 5 **EXCELLENT:** A precise awareness of audience is effectively sustained. Development of topic or function is clearly focused and effective. Significant information is presented, and this information is enhanced by precise and appropriate details that effectively fulfil the purpose.
 - 4 **PROFICIENT:** Awareness of audience is clearly sustained. Development of topic or function is generally effective. Significant information is presented, and this information is substantiated by appropriate details that efficiently fulfil the purpose.
 - 3 **SATISFACTORY:** Awareness of audience is generally sustained. Development of topic or function is adequate. Sufficient information is presented, and this information is supported by enough detail to fulfil the purpose.
 - 2 **LIMITED:** Awareness of audience is evident but is not sustained. Development of topic or function is vaguely focused and ineffective. Essential information may be missing. Supporting details are scant, insignificant, and/or irrelevant. The purpose is only partially fulfilled.
 - 1 **POOR:** Little awareness of audience is evident. Development of topic or function, if present, is obscure. Essential information and supporting details are inappropriate or lacking. The purpose is not fulfilled.
- INS INSUFFICIENT:** The marker can discern no evidence of an attempt to fulfil the assignment, or the writing is so deficient in length that it is not possible to assess thought and detail.

Section II: Functional Writing - Scoring Guide (continued)

Writing Skills (curriculum concepts 2, 3, 4, 5)

When marking Writing Skills, the marker should consider Matters of Choice AND Matters of Convention.

PROPORTION OF ERROR TO COMPLEXITY AND LENGTH OF RESPONSE MUST ALSO BE CONSIDERED.

- 5 EXCELLENT: The selection and use of words and structures are effective. Minor errors in mechanics and grammar do not reduce the clarity of communication.
- 4 PROFICIENT: The selection and use of words and structures are usually effective. Very rarely do minor errors in mechanics and grammar reduce the clarity of communication.
- 3 SATISFACTORY: The selection and use of words and structures are generally effective. Errors in mechanics and grammar may occasionally reduce the clarity of communication.
- 2 LIMITED: The selection and use of words and structures may be ineffective. Errors in mechanics and grammar reduce the clarity of communication.
- 1 POOR: The selection and use of words and structures are frequently ineffective. Errors in mechanics and grammar severely reduce the clarity of communication.

Section III: Response to Visual Communication - Scoring Guide

Thought and Detail (curriculum concepts 1, 3, 4, 5, 13, 14, 16, 17)

When marking Thought and Detail, the marker should consider how effectively the details selected from the photograph and the discussion of these details contribute to a plausible and consistent interpretation of the photograph.

- 5 **EXCELLENT:** Interpretation of the photograph is insightful and is in the form of an effective generalized idea or theme. Specific details used for support are purposefully chosen and enhance clarity.
- 4 **PROFICIENT:** Interpretation of the photograph is well considered and is in the form of a generalized idea or theme. Specific details used for support are well defined and accurate.
- 3 **SATISFACTORY:** Interpretation of the photograph is conventional and may be in the form of a maxim or moral. Details used for support are clear but tend to be generalized.
- 2 **LIMITED:** Interpretation of the photograph is vague and uncertain and/or concentrated on a particular detail rather than on the photograph as a whole. Details used for support are inappropriate and/or unclear.
- 1 **POOR:** Interpretation of the photograph is inappropriate or incomprehensible. Details are irrelevant, inaccurate, or absent.
- INS INSUFFICIENT:** The marker can discern no evidence of an attempt to fulfil the assignment, or the writing is so deficient in length that it is not possible to assess thought and detail.

Section III: Response to Visual Communication - Scoring Guide

Organization (curriculum concepts 2, 3, 4)

When marking Organization, the marker should consider how effectively the writing demonstrates

- unified and consistent development
- clear and coherent order

- 5 **EXCELLENT:** A controlling idea is clear and is successfully sustained. Ideas are developed by sentences and paragraphs that flow smoothly and coherently.
- 4 **PROFICIENT:** A controlling idea is clear and is usually sustained. Ideas are developed by sentences and paragraphs that are coherently related.
- 3 **SATISFACTORY:** A controlling idea is mechanically maintained. Sentences and paragraphs are generally related, but coherence may falter on occasion.
- 2 **LIMITED:** A controlling idea may be lacking or not maintained. Relationships between sentences and between paragraphs are frequently unclear.
- 1 **POOR:** A controlling idea is lacking. Sentences and paragraphs are not coherently related.

Writing Skills (curriculum concepts 2, 3, 4, 5)

When marking Writing Skills, the marker should consider Matters of Choice AND Matters of Convention.

PROPORTION OF ERROR TO COMPLEXITY AND LENGTH OF RESPONSE MUST ALSO BE CONSIDERED.

- 5 **EXCELLENT:** The selection and use of words and structures are effective. Minor errors in mechanics and grammar do not reduce the clarity of communication.
- 4 **PROFICIENT:** The selection and use of words and structures are usually effective. Very rarely do minor errors in mechanics and grammar reduce the clarity of communication.
- 3 **SATISFACTORY:** The selection and use of words and structures are generally effective. Errors in mechanics and grammar may occasionally reduce the clarity of communication.
- 2 **LIMITED:** The selection and use of words and structures may be ineffective. Errors in mechanics and grammar reduce the clarity of communication.
- 1 **POOR:** The selection and use of words and structures are frequently ineffective. Errors in mechanics and grammar severely reduce the clarity of communication.

Appendix B

Equations For Calculation Of Variance Components

EQUATIONS FOR CALCULATION OF VARIANCE COMPONENTS

<u>Source of variation</u>	<u>Equation</u>
rater-scale interaction	$\sigma^2_{ji} = (EMS_{ji} - \sigma^2_{nij,e}) / n_n$
examinee-scale interaction	$\sigma^2_{ni} = (EMS_{ni} - \sigma^2_{nij,e}) / n_j$
examinee-rater interaction	$\sigma^2_{nj} = (EMS_{nj} - \sigma^2_{nij,e}) / n_i$
scale	$\sigma^2_i = (EMS_i - \sigma^2_{nij,e} - n_n \sigma^2_{ij} - n_j \sigma^2_{ni}) / n_n n_j$
rater	$\sigma^2_j = (EMS_j - \sigma^2_{nij,e} - n_n \sigma^2_{ij} - n_i \sigma^2_{nj}) / n_n n_i$
examinee	$\sigma^2_n = (EMS_n - \sigma^2_{nij,e} - n_i \sigma^2_{nj} - n_j \sigma^2_{ni}) / n_j n_i$

Appendix C

Derivation of the Multifaceted Rasch Model

This text is taken essentially unaltered from Chapter Five of *Many-Facet Rasch Measurement* (Linacre, 1989). As such, equation numbers and figure numbers, remain as in Chapter Five.

The particular model to be derived here is applicable to a three-faceted test in which each judge of a panel of judges awards a rating to each examinee on each item.

Consider the performance of two examinees, E_m and E_n , rated by a Judge J_j on replications of the item A_i . In whatever way the ratings were originally recorded, they have been recoded into $K+1$ categories ordinarily numbered from 0 to K , with each higher numbered category representing a higher level of perceived performance, and with each category having a non - zero probability of occurrence.

The administrations of the numerous replications of item A_i is the "test". The performance levels of examinees E_n and E_m can be compared by their relative frequencies of being rated in the various categories of the rating scale. Part of their performance can be summarized by a 2x2 cross-tabulation table of counts of rating in categories k and h of the rating scale, chosen so that category k is numerically greater than category h and represents a higher performance level. This category is depicted in Fig. 10.

		Examinee E_n	
Categories		k	h
Examinee E_m	k	F_{kk}	F_{hk}
	h	F_{kh}	F_{hh}

Figure 10. Frequency distribution of judge-awarded ratings. F_{kh} represents the count of the number of times that examinee E_n is awarded rating k , when examinee E_m is rated an h , by judge J_j across numerous replications of item A_i , where $k > h$.

When both examinees are given the same rating, which occurs F_{kk} times for a rating of k , and F_{hh} times for a rating of h , their performance levels are indistinguishable. When the examinees are rated differently, which occurs F_{kh} and F_{hk} times, the examinee with the greater frequency of ratings in category k , the higher category, is perceived to have the higher ability. In comparing performance levels, we intend that the numeric result be independent of the number of replications. Thus, if the test were to be repeated again, and were of the same length, we would expect to get approximately the same result. Moreover, if the two tests were then to be concatenated, we would again expect to obtain about the same result. The division of the two frequencies, F_{kh} and F_{hk} , is compatible with this expectation because we expect this ratio to be about the same when the test is repeated, and also when the two tests are concatenated. Consequently, the comparative levels of performance of examinees E_n and E_m can be identified by the ratio, F_{kh} / F_{hk} .

$$\frac{PL(E_n)}{PL(E_m)} \approx \frac{F_{kh}}{F_{hk}}, \quad 5.1$$

where $PL(E_n)$ is the performance level of E_n , and
 $PL(E_m)$ is the performance level of E_m ,

The ratio of empirically observed frequencies, F_{kh} / F_{hk} , is an approximation, which is never exact, to the ratio of probabilities, P_{kh} / P_{hk} , where P_{kh} is the probability of examinee, E_n , being given a rating of k , when examinee, E_m , is being given a rating of h where $k > h$, and P_{hk} is the probability of examinee, E_n , being given a rating of h , when examinee, E_m , is being given a rating of k . This unobservable ratio, P_{kh} / P_{hk} is defined to be the ratio of the examinees' performances.

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{kh}}{P_{hk}}, \quad 5.2$$

Now, for objectivity, the ratings given examinees E_n and E_m must be independently awarded by the judge. Consequently,

$$P_{kh} = P_{nij k} \times P_{mij k} \quad 5.3$$

and

$$P_{hk} = P_{nij h} \times P_{mij h} \quad 5.4$$

where $P_{nij k}$ is the probability of examinee E_n being given a rating of k on item A_i by judge J_j . $P_{nij h}$, $P_{mij k}$, $P_{mij h}$, are similarly defined. Then

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{kh}}{P_{hk}} = \frac{P_{nij k}}{P_{nij h}} \times \frac{P_{mij k}}{P_{mij h}} \quad 5.5$$

Furthermore, also for objectivity, the relative performance of examinees E_n and E_m must be independent of which particular item is used to compare them. Thus, though performance levels are initially defined in terms of any conceptually equivalent items A_i' . That is,

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{nij k}}{P_{nij h}} \times \frac{P_{mij h}}{P_{mij k}} = \frac{P_{n i' j k}}{P_{n i' j h}} \times \frac{P_{m i' j h}}{P_{m i' j k}} \quad 5.6$$

then

$$\frac{P_{nij k}}{P_{nij h}} = \frac{P_{mij k}}{P_{mij h}} \times \frac{P_{n i' j k}}{P_{n i' j h}} \times \frac{P_{m i' j h}}{P_{m i' j k}} \quad 5.7$$

For objectivity, this ratio of the probabilities of examinees E_n being rated in categories k and h must be independent of whichever examinee E_m is used in the comparison. Now consider examinee E_0 with performance level at the local origin of the ability sub-scale. Similarly the ratio must be independent of whichever item A_i' is used for the comparison. Thus it must also hold for item A_0 chosen to have difficulty at the local origin of the item sub-scale.

$$\frac{P_{nij k}}{P_{nij h}} = \frac{P_{0ij k}}{P_{0ij h}} \times \frac{P_{n0 j k}}{P_{n0 j h}} \times \frac{P_{00 j h}}{P_{00 j k}} \quad 5.8$$

If, instead of comparing performance levels by means of items A_i and A_i' , we compare performance levels by means of the ratings given by judges J_j and J_j' over numerous replications of item A_i , then again we expect the relative performance levels to be maintained.

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{nij\bar{k}}}{P_{nijh}} \times \frac{P_{mijh}}{P_{mijk}} = \frac{P_{nij\bar{k}}}{P_{nijh}} \times \frac{P_{mij'h}}{P_{mij'k}} \quad 5.9$$

so that

$$\frac{P_{nij\bar{k}}}{P_{nijh}} = \frac{P_{mijk}}{P_{mijh}} \times \frac{P_{nij\bar{k}}}{P_{nijh}} \times \frac{P_{mij'h}}{P_{mij'k}} \quad 5.10$$

Again this must be true if judge J_j' is chosen to be judge J_0 with severity at the local origin of the severity scale, and examinee E_m is examinee E_0 , and when item A_i is replaced by item A_0 . Therefore,

$$\frac{P_{n0\bar{k}}}{P_{n0jh}} = \frac{P_{00\bar{k}}}{P_{00jh}} \times \frac{P_{n00\bar{k}}}{P_{n00h}} \times \frac{P_{000h}}{P_{000k}} \quad 5.11$$

Furthermore, for objectivity, the relative severity level (SL) of judges J_j and J_j' must be maintained whether the judging takes place over numerous replications of the administration of either item A_i or item A_i' to the same examinee E_n .

$$\frac{SL(J_j')}{SL(J_j)} = \frac{P_{nij\bar{k}}}{P_{nijh}} \times \frac{P_{n'ijh}}{P_{n'ij\bar{k}}} = \frac{P_{nij\bar{k}}}{P_{nijh}} \times \frac{P_{n'ij'h}}{P_{n'ij'k}} \quad 5.12$$

then

$$\frac{P_{nij\bar{k}}}{P_{nijh}} = \frac{P_{n'ij\bar{k}}}{P_{n'ijh}} \times \frac{P_{n'ij'h}}{P_{n'ij'k}} \times \frac{P_{n'ij'h}}{P_{n'ij'k}} \quad 5.13$$

Again this must be true if judge J_j' is judge J_0 chosen at the origin of the severity scale, and examinee E_n is examinee E_0 , and item A_i' is item A_0 .

$$\frac{P_{ijk}}{P_{ijh}} = \frac{P_{00jk}}{P_{00jh}} \times \frac{P_{0i0k}}{P_{0i0h}} \times \frac{P_{000k}}{P_{000h}} \quad 5.14$$

Substituting equations 5.14 and 5.11 in 5.8, and simplifying,

$$\frac{P_{ijk}}{P_{ijh}} = \frac{P_{n00k}}{P_{n00h}} \times \frac{P_{0i0k}}{P_{0i0h}} \times \frac{P_{00jh}}{P_{00jk}} \times \left(\frac{P_{000h}}{P_{000k}} \right)^2 \quad 5.15$$

which gives a general form in which each term is an expression of the relationship between a component of a facet and the local origin of a sub-scale, in the context of a particular pair of categories.

As the more general circumstances of a rating scale are to be considered, we do not wish the comparison of the abilities of E_n and E_m to depend on which particular pair of categories of the rating scale that are used. Beginning again with equation 5.5

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{kh}}{P_{hk}} = \frac{P_{ijk}}{P_{nijh}} \times \frac{P_{mijk}}{P_{mijh}} \quad 5.5$$

We wish to generalize this equation to any pair of categories. The rating scale, categories, are not independent but structured. In order to determine the structure in an objective manner, we require that performance levels are invariant when they are compared using any pair of adjacent categories in ascending order. This is the only possible objective structuring, since defining invariance not over adjacent categories, but over some pairing of non adjacent, results in a contradiction or indeterminacy in the rating scale structure. Thus if a rating scale has 3 categories and performance levels are to be invariant only when the top and bottom categories are used for the comparison in equation 5.15, then performance levels on the middle category are indeterminate, and so not objective.

Invariance in relative performance when categories chosen such that k is one greater than h , and also k' is chosen one greater than h' , yields

$$\frac{PL(E_n)}{PL(E_m)} = \frac{P_{nijk}}{P_{nijh}} \times \frac{P_{mijh}}{P_{mijk}} = \frac{P_{nijk'}}{P_{nijh'}} \times \frac{P_{mijh'}}{P_{mijk'}} \quad 5.19$$

so that

$$\frac{P_{nijk}}{P_{nijh}} = \frac{P_{nijk'}}{P_{nijh'}} \times \frac{P_{mijk}}{P_{mijh}} \times \frac{P_{mijh}}{P_{mijk'}} \quad 5.20$$

Since the result is to be generalizable, a substitution is made with examinee E_0 replacing E_m , item A_0 for item A_i , and judge J_0 for judge J_j , where E_0 is an examinee with ability 0, A_0 is an item of difficulty 0, and J_0 is a judge of severity 0, that is, the examinee, item and judge all possess the mean amount of their respective properties. A series of repeated substitutions and reorderings leads to the equation

$$\frac{P_{n00k}}{P_{n00h}} = \frac{P_{n00k'}}{P_{n00h'}} \times \frac{P_{000k}}{P_{000h}} \times \frac{P_{000k'}}{P_{000h'}} \quad 5.20$$

Reordering the terms,

$$\frac{P_{n00k}}{P_{n00h}} = \left(\frac{P_{n00k'}}{P_{n00h'}} \times \frac{P_{000k'}}{P_{000h'}} \right) \times \frac{P_{000k}}{P_{000h}} \quad 5.20$$

The two terms in parentheses are invariant over changes in choice of pairs of categories and so are independent of the local structure of the rating scale, but they are not independent of the choice of object, so we can accordingly write them as P_{n00} , so that

$$\frac{P_{n00k}}{P_{n00h}} = P_{n00} \times \frac{P_{000k}}{P_{000h}} \quad 5.20$$

Similar equations hold for P_{0i0k} / P_{0i0h} and P_{0i0k} / P_{0i0h} , so that, substituting into 5.15,

$$\frac{P_{nijk}}{P_{nijh}} = P_{n00} \times P_{0i0} \times P_{00j} \times \frac{P_{000k}}{P_{000h}}$$

This is the equation in which the ratio of probabilities of particular outcomes is the product of terms that depend only on a single component and the local origin of its sub-scale, combined with a term dependent on the pair of categories used for comparison.

The natural logarithm of this equation, followed by the following definitions

$B_n = \ln(P_{n00})$, the ability of examinee E_n ,

$D_i = -\ln(P_{0i0})$, the difficulty of item A_i ,

$C_j = -\ln(P_{00j})$, the severity of judge J_j

$F_k = -\ln\left(\frac{P_{000k}}{P_{000h}}\right)$, the difficulty of the step from category $k-1$ to category k .

yields

$$\ln\left(\frac{P_{nijk}}{P_{nij(k-1)j}}\right) = B_n - D_i - C_j - F_k$$

which is the equation for the multifaceted Rasch model for three facets and a rating scale.

Appendix D

Weighted Total Mean Score and Unweighted Total Mean Score by Rater

Weighted Total Mean Score and Unweighted Total Mean Score by Rater

Rater Number	Weighted Mean Score	Unweighted Mean Score	%Weighted Mean Score	%Unweighted Mean Score	Correlation
1	28.11	25.11	56.22	55.80	.98
2	30.68	27.44	61.36	60.98	.98
3	29.58	26.24	59.16	58.31	.99
4	30.33	26.73	60.66	59.40	.99
5	27.40	24.27	54.80	53.93	.98
6	30.23	27.00	60.46	60.00	.99
7	30.44	26.80	60.88	59.56	.98
8	29.57	26.84	59.14	59.64	.98
9	29.26	26.41	58.52	58.69	.97
10	27.85	25.12	55.70	55.82	.98
11	29.70	26.78	59.40	59.51	.99
12	32.60	28.82	65.20	64.04	.98
13	30.81	27.48	61.62	61.07	.99
14	29.22	26.01	58.44	57.80	.98
15	30.05	26.98	60.10	59.96	.98
16	28.95	25.91	57.90	57.58	.99
17	29.51	26.12	59.02	58.04	.99
18	29.20	26.02	58.40	57.82	.98
19	29.95	26.19	59.90	58.20	.98
20	29.20	26.01	58.40	57.80	.98
21	28.54	25.12	57.08	55.82	.98
22	33.36	30.02	66.72	66.71	.99
23	29.03	26.13	58.06	58.07	.98
24	31.03	27.76	62.06	61.69	.99
25	30.98	27.50	61.96	61.11	.99
26	29.81	26.84	59.62	59.64	.99
27	32.84	29.31	65.68	65.13	.99
28	30.78	27.48	61.56	61.07	.99
29	30.80	27.23	61.60	60.51	.99
30	28.66	25.52	57.32	56.71	.99
31	28.17	25.29	56.34	56.20	.99
32	28.29	25.45	56.58	56.56	.99
33	29.28	25.84	58.56	57.42	.99
34	30.75	27.58	61.50	61.29	.99
35	28.70	25.77	57.40	57.27	.99
36	26.80	23.73	53.60	52.73	.98
37	29.20	26.04	58.40	57.87	.99
38	29.42	26.25	58.84	58.33	.98
39	29.96	26.83	59.92	59.62	.98
40	32.49	28.56	64.98	63.47	.98
41	29.47	26.24	58.94	58.31	.99
42	33.11	29.76	66.22	66.13	.98
43	29.11	26.30	58.22	58.44	.99

Rater Number	Weighted Mean Score	Unweighted Mean Score	%Weighted Mean Score	%Unweighted Mean Score	Correlation
44	30.67	27.24	61.34	60.53	.99
45	30.53	27.27	61.06	60.60	.99
46	33.19	29.36	66.38	65.24	.98
47	29.39	26.42	58.78	58.71	.99
48	29.56	26.32	59.12	58.49	.99
49	29.30	26.22	58.60	58.27	.99
50	28.19	25.07	56.38	55.71	.99
51	30.20	26.84	60.40	59.64	.98
52	27.79	24.51	55.58	54.47	.99
53	32.27	28.67	64.54	63.71	.98
54	29.26	26.02	58.52	57.82	.99
55	30.93	27.14	61.86	60.31	.99
56	28.74	25.72	57.48	57.16	.99
57	27.11	24.33	54.22	54.07	.98
58	25.43	22.88	50.86	50.84	.98
59	27.15	24.17	54.30	53.71	.98
60	31.21	27.65	62.42	61.44	.98
61	31.51	28.10	63.02	62.44	.98
62	31.19	27.88	62.38	61.96	.99
63	30.22	27.09	60.44	60.20	.98
64	32.23	28.58	64.46	63.51	.99
65	28.25	25.15	56.50	55.89	.97
66	29.49	26.90	58.98	59.78	.98
67	29.49	26.24	58.98	58.31	.98
68	27.29	24.37	54.58	54.16	.98
69	28.56	25.75	57.12	57.22	.98
70	28.37	25.54	56.74	56.76	.99

Appendix E

Demonstration Of The Feasibility Of The Proposed Generalizability Sampling Design
And Subsequent Analysis

ADDENDUM TO DISSERTATION PROPOSAL

DEMONSTRATION OF THE FEASIBILITY OF THE PROPOSED
GENERALIZABILITY SAMPLING DESIGN AND SUBSEQUENT ANALYSIS

The Sample

The English 33 data were sorted by the third marker identification number. Given that the number of possible marker triplets far exceeds the number of bundles, the use of marker numbers allowed the specification of combinations of three markers that actually did mark a common bundle of papers. Marker identification numbers were considered likely to produce a random sample of the student population as student identification numbers or student record numbers would group students according to school and region, whereas marker identification numbers bear no relation to marker behaviour or any student characteristic. The third marker was picked as the sort variable by chance.

A sample was drawn so that 100 intact bundles of six were chosen. This sample contained all intact bundles for markers 301 to 313 plus sufficient bundles from rater 314 to form the 100 bundles for this sample. This sample was grouped in units of 20 intact bundles of six. Groups of twenty bundles were chosen to minimize researcher fatigue and researcher error. The group composition characteristics are given in Table A1.

Table A1
The Generalizability Group Composition Characteristics

Group Number	Bundle				Comments
	Number	Size	Quantity	Total N	
1	01-20	6	20	120	
2	21-40	6	20	240	
3	41-60	6	20	360	
4	61-80	6	20	480	
5	81-100	6	20	600	
6	F1-F25	5	25	735	2 - 6's, 1 - 7, 1 - 11, N=135
7	A1-A41	6	41	246	all Marker 353; no overlap
8	T1-T11	12	11	173	bundle size 8 - 12; mainly 12

In addition to the first five groups, three other subsamples were drawn. While Groups 1 to Group 5 comprise all of the intact bundles of six, Group 6 consists of all the bundles of five, chosen from the same sample group of markers as was Group 1 to Group

5. These bundles are coded in Table A1 as F1 to F25. As reported in Chapter Two, five is the most common number of papers in any broken bundle. Group 6 was chosen in order to obtain a more nearly complete subsample of markings from markers 301 to 314.

Two other subsamples, Group 7 and Group 8 were chosen in order to compare the results of different sampling methods. Group 7 consists of all intact bundles in which marker 353, the most prolific marker, was one of the three markers. These bundles are coded in Table A1 as A1 to A41. There is a no overlap with the main sample as papers that were in the main sample were excluded from this sample. Group 8 consists of groups for third markers numbered 301 to 314 in which units are formed from groups of students who share two common markers; this sampling giving larger units for generalizability analysis. These bundles are coded in Table A1 as T1 to T11.

The Analysis

First Analysis

Each bundle was analyzed according to a fully crossed design, examinees by markers by scales, $n \times j \times i$, using the computer program BMDP 8V. The percentage of total variance was calculated for each component within each bundle. The percentages for the variance components determined for each bundle were aggregated into percentage variance components for the groups described in Table A1. The aggregated variance components were then compared for stability of estimates both within the groups, Group 1 to Group 6, and between the total of Group 1 to Group 6, and Group 7, and Group 8. The results for these two subsamples and the total results with and without these subsamples are given in Table A2.

Table A2
Percent Variance Components for the Subsamples

Group	%N	%J	%I	%NJ	%NI	%JI	%NJI,E
1	20.65	6.32	1.33	14.01	13.58	4.04	40.08
2	28.71	7.72	2.05	13.03	13.02	3.20	32.27
3	25.21	5.02	1.34	10.13	20.90	4.20	33.20
4	29.01	5.72	1.69	14.02	14.05	2.87	32.63
5	29.02	4.02	- 0.55	12.25	16.98	4.84	33.45
6	24.56	4.17	0.31	12.98	16.92	4.59	36.48
7	32.12	3.62	0.46	11.10	14.39	4.61	33.70
8	24.31	3.88	- 0.94	19.71	16.04	1.76	34.46
Total 1-6	26.21	5.38	0.93	12.72	16.00	4.02	34.72
Total 1-7	27.64	4.96	0.82	12.33	15.61	4.17	34.48
Total 1-8	27.32	4.85	0.73	13.03	15.65	3.93	34.47

Table A3
Final Sample Percent Variance Components for the Subsamples

Group	%N	%J	%I	%NJ	%NI	%JI	%NJI,E
1	20.651	6.317	1.327	14.006	13.579	4.039	40.080
2	28.712	7.636	2.047	13.033	13.020	3.208	32.343
3	25.169	5.016	1.305	10.475	20.714	4.203	33.118
4	29.722	5.415	1.505	13.719	14.883	2.954	31.801
5	28.409	3.644	-0.596	12.387	16.718	4.998	34.438
6	27.501	2.488	0.003	10.027	21.058	3.751	35.172
7	34.105	2.946	-0.052	10.754	15.031	5.158	32.058
8	32.804	3.786	2.990	10.306	15.259	1.908	32.947
9	24.235	9.278	0.679	11.495	17.487	3.314	33.512
Total 1-9	27.986	5.100	1.029	11.805	16.399	3.733	33.948

Group	%N	%J	%I	%NJ	%NI	%JI	%NJI,E
Total 1-9	27.986	5.100	1.029	11.805	16.399	3.733	33.948

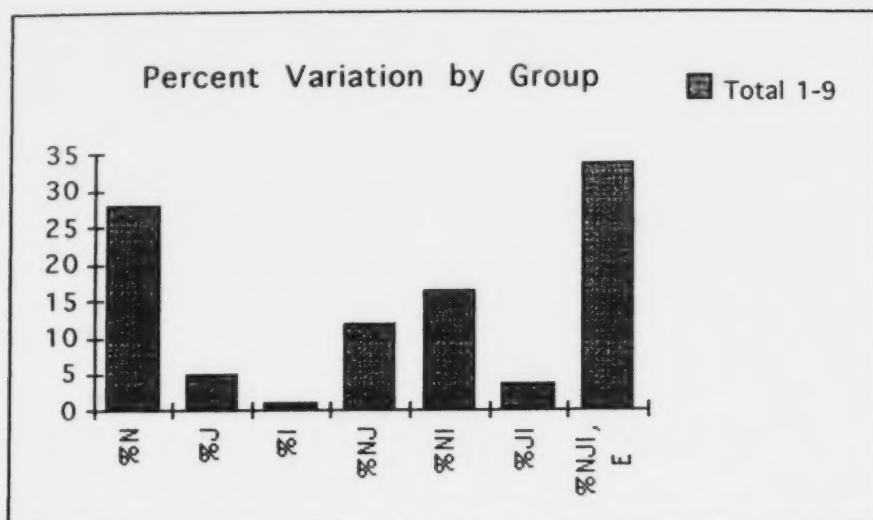


Figure A1
Bar graph of Percent variance components for the subsamples.

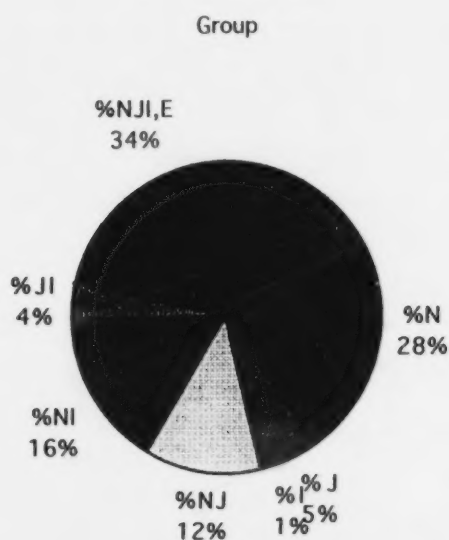


Figure A2
Circle graph of Percent variance components for the subsamples.

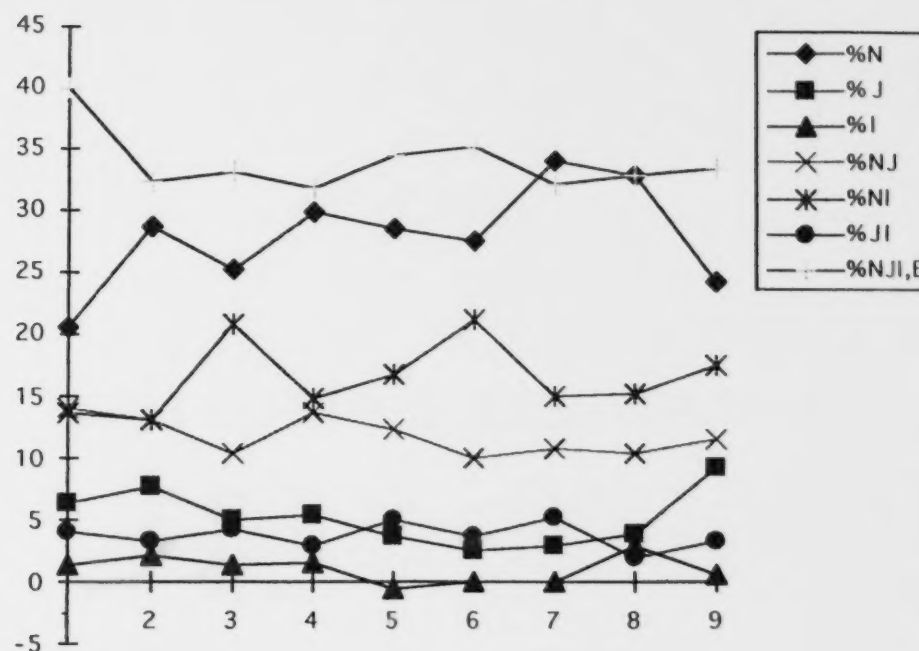


Figure A3

Circle graph of Percent variance components for the subsamples.

Considerable variation occurs among the means of the groups of 20. This is expected given the small number of subjects per group. In spite of the large variation, several features are already apparent. The NJI,error component and the person component are largest by far, hovering around 35% and 27% respectively. The interactions, person - judge, and person - item, account for approximately 13% and 16% respectively. The remaining components, judge, item, and the judge - item interaction all are low, 6%, 1%, and 4% respectively. Both this ranking of components, and the approximate magnitude are apparent even after the first group of 20. The order and magnitude are stable throughout the various groups chosen.

Examination of variance components of Group 7, all intact bundles marked by marker 353, reveal no differences that suggest that analysis of a different part of the sample will lead to different conclusions about the various facets. Likewise, examination of variance components of Group 8, comprised of larger units with two markers in common,

reveal no differences that suggest that analysis of a different unit within the sample will lead to different conclusions about the various facets.

These results are presented in graphical form in Figure A4.



Figure A4.
Percent variance components for the subsamples.

Given the lack of distinct differences in estimates of variance components for Group 7 or Group 8 and the original six groups, only analyses involving the first six groups will be pursued.

Nature Of Individual Bundle Variance Components

Given that the variance estimates were produced with samples consisting of five or six cases, questions may nevertheless be raised about the nature of estimates produced by

the analyses of the bundles. Table A3 contains information pertaining to the distribution of these variance estimates.

Table A3
Individual Bundle Variance Components Characteristics

Component	Mean	Median	SD.	Skew	Kurtosis
%N	26.22	25.31	17.14	0.07	-0.66
%J	5.38	2.62	7.92	1.34	1.59
%I	0.93	0.00	4.91	0.68	0.33
%NJ	12.72	11.58	8.47	0.83	1.53
%NI	16.00	13.36	11.30	1.16	1.43
%JI	4.02	3.06	4.63	1.24	1.73
%NJI,E	34.72	34.10	12.54	0.50	0.72

The median is relatively different than the mean in the judge facet, the examinee - item, and the judge - item cases. Skew and kurtosis values for these facets are larger than what would be expected for a normal distribution. The judge facet values indicate that some bundles contain a judge who differs greatly from his/her fellow judges. The judge - item interaction suggests that some judges behave quite differently across some scales. The indication of some examinees behaving differently across items, as the examinee - item interaction suggests, is not surprising as some examinees will opt not to attempt a section of the examination but will perform reasonably on other sections.

Standard Deviation versus Standard Error. As both standard deviation and standard error are given, explanation of their meaning as they relate to this study follows. The standard deviation relates to the amount of spread among estimates of the variance component that would occur if other bundles of examinees were analyzed. It is the standard error of the mean for a sample of one bundle. The investigation of variation in estimates among very small n generalizability studies is not the purpose of this study. Therefore, the standard deviation is not of interest in this study. The standard error relates to the amount of spread among estimates of the mean of the variance component that would occur in samples consisting of many bundles. The goal of this study is produce estimates

of variance components that reflect the population being studied. It is the precision of these estimates of the variance that is of interest in this study. The standard error provides a measure of this precision.

The percentage variance components with their 95% confidence intervals for the sample consisting of Group 1 to Group 6 are given in Table A4 and Figure A5.

Table 4A

Percentage Variance Components with Confidence Intervals

Estimate	Mean	SD	SE.	95% Confidence Interval for Mean	
%N	26.22	17.14	1.51	23.23	29.20
%J	5.38	7.92	0.69	4.00	6.76
%I	0.93	4.91	0.43	0.08	1.79
%	12.72	8.47	0.75	11.24	14.19
%NI	16.00	11.30	0.99	14.03	17.97
%JI	4.02	4.63	0.41	3.22	4.83
%NJI,E	34.73	12.54	1.10	32.54	36.91

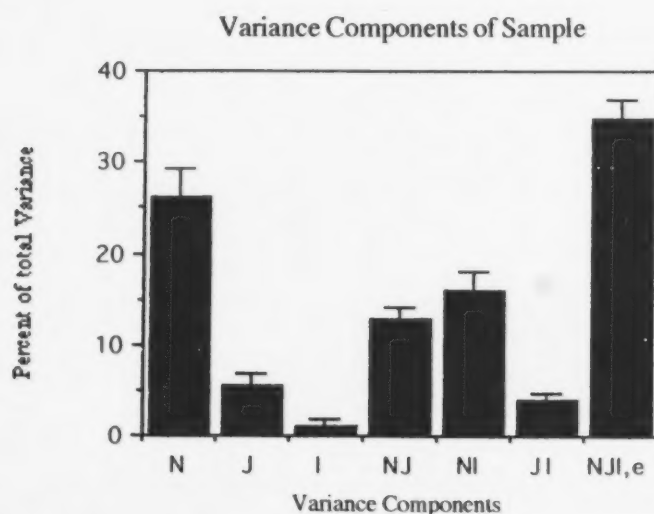


Figure A5

Bar graph of percentage variance components.

As both Table A4 and Figure A5 both clearly suggest, the method of sampling proposed for the generalizability study will lead to estimates for variance components that are both stable and sufficiently precise such that the results of a generalizability can be readily interpreted. Nevertheless, a variation of the preliminary analysis is given to further strengthen this claim.

Second Analysis

The second analysis differs from the first in two respects, apart from the use of only the first six groups. First, variance component values themselves rather than percentages will be aggregated. Second, the results for the groups of 20 will be displayed in a cumulative fashion, rather than separate groups of twenty, the second group will now consist of an aggregation of the variance components of the first 40 bundles rather than being comprised of bundle 21 through bundle 40. The results of this analysis are first presented in tabular form in Table A5.

Table A5
Cumulative Variance Components for the Groups

	N	J	I	NJ	NI	JI	NJI,E
1	0.159	0.043	0.008	0.090	0.090	0.028	0.253
1-2	0.192	0.047	0.009	0.092	0.094	0.024	0.231
1-3	0.187	0.042	0.009	0.080	0.112	0.025	0.222
1-4	0.200	0.040	0.011	0.087	0.111	0.023	0.224
1-5	0.202	0.038	0.010	0.086	0.113	0.025	0.221
1-6	0.207	0.035	0.007	0.085	0.119	0.026	0.221
1-6	29.6%	5.0%	1.3%	12.1%	17.0%	3.7%	31.5%

In spite of the limited number of data points, several features are apparent. As before, the NJI,error component and the person component are the largest by far, at 32% and 30% respectively. Interactions, person and judge, and person and item, account for 12% and 17% respectively. The remaining components, judge, item, and the judge - item interaction all are low, 5%, 1%, and 4% respectively. The results are presented in graphical form in Figure A6.

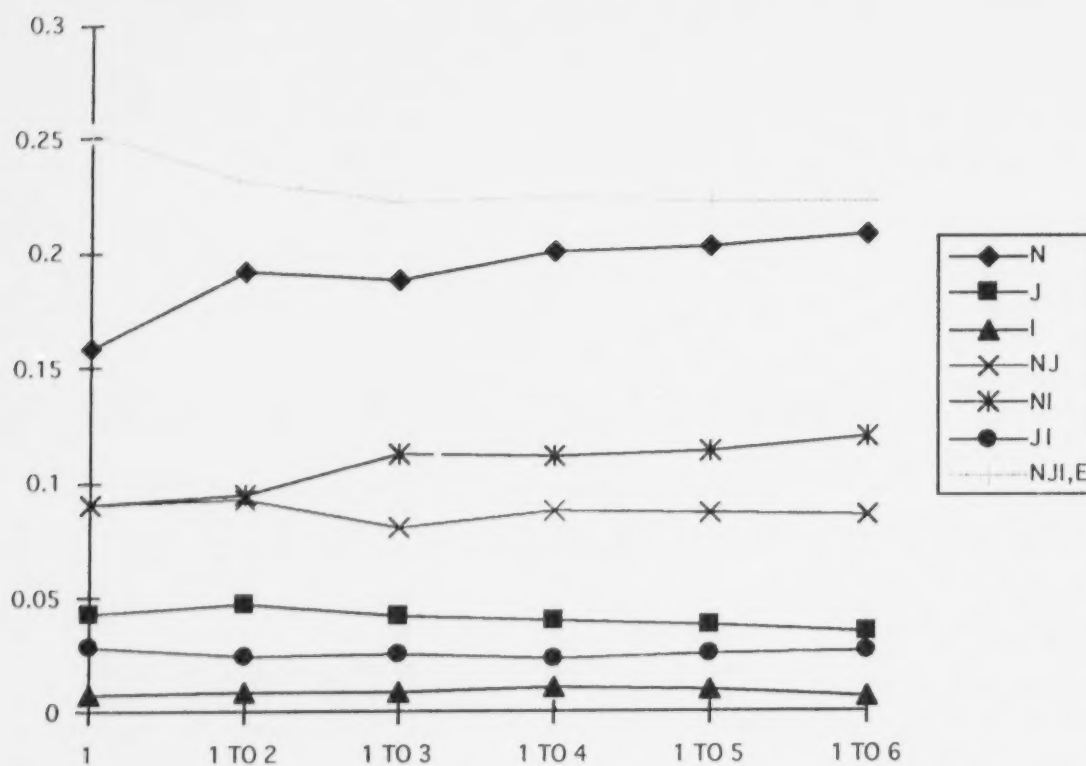


Figure A6.
Cumulative variance components for the groups.

The cumulative estimates appear to be approaching asymptotic values yet this can only be hinted at given the limited number of points.

As has been done with the first analysis, the characteristics of the cumulative group are presented with confidence intervals in Table A7. In this table, actual variance values, rather than percentages, are given.

Table A7
Variance Components with Confidence Intervals

Estimate	Mean	SD	SE.	95% Confidence Interval for Mean	
N	0.207	0.180	0.016	0.175	0.239
J	0.035	0.054	0.005	0.025	0.045
I	0.007	0.035	0.003	0.001	0.013
NJ	0.085	0.064	0.006	0.073	0.097
NS	0.119	0.115	0.010	0.099	0.139
JI	0.026	0.032	0.003	0.020	0.032
NJI,E	0.221	0.062	0.006	0.209	0.233

Again as the results displayed in Figure A7, and presented in Table A7, indicate stable, interpretable variance estimates were obtained. Due to the great similarity to the results of the first analysis, the bar graph of the variance components will not be repeated.

Summary and Conclusions

Variance estimates of individual bundles vary widely as expected but produce mean values that are consistent across subsamples. The means of the subsamples illustrate that the method of obtaining population estimates by aggregating estimates produced by analysis of bundles is a feasible method for analyzing data that cannot otherwise be analyzed because of the characteristics of the data matrix.